Exploiting the speech–gesture link to capture fine-grained prosodic prominence impressions and listening strategies

Article *in* Journal of Phonetics · July 2019

DOI: 10.1016/j.wocn.2019.07.001

CITATIONS
6

READS
262

3 authors:

Petra Wagner
Bielefeld University
176 PUBLICATIONS **1,370** CITATIONS
SEE PROFILE

Aleksandra Ćwiek
Centre for General Linguistics
8 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Barbara Samlowski
Amazon Germany
10 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Exploiting the speech-gesture link to capture fine-grained prosodic prominence impressions and listening strategies

Petra Wagner[1], Aleksandra Ćwiek[1,2], Barbara Samlowski[1,3]

Bielefeld University (1), Leibniz-Centre General Linguistics (2), Amazon Development Center Germany (3)

Abstract

In this paper, we explore the possibility to gather perceptual impressions of prosodic prominence by exploiting the strong prosody-gesture link, i.e., by having listeners transform a perceptual impression into a motor movement, namely drumming, for two domains of prominence: word-level and syllable-level. A feasibility study reveals that such a procedure is indeed easily and speedily mastered by naïve listeners, but more difficult for word-level prominences. We furthermore examine whether "drummed" annotations are comparable to those gathered with more established annotation protocols based on cumulative naïve impressions and fine-grained expert ratings. These comparisons reveal high correspondences across all prominence annotation protocols, thus corroborating the general usefulness of the gestural approach. The analyses also reveal that all annotation protocols are strongly driven by structural linguistic considerations. We then use Random Forest Models to investigate the relative impact of signal and structural cues to prominence annotations. We find that expert ratings of prosodic prominence are guided comparatively more by structural concerns than those of naïve annotators, that word-level annotations are influenced more by structural linguistic cues than syllable-level ones, and that "drummed" annotations are driven least by structural cues. Lastly, we isolate two main listener strategies among our group of "drummers", namely those integrating structural and signal cues to prominence, and those being guided predominantly by signal cues.

Keywords: prominence, prosody, annotation, signal correlates, structural correlates, prominence domains, gestures

1. Introduction

To this day, no standard of the concept of prosodic prominence or its adequate measurement exists. Rather, prominence seems to serve as a loose umbrella term for a set of related phonological and phonetic concepts such as lexical stress, pitch accent, sentence stress, prosodic focus, rhythmic alternation, metrical grid height, paralinguistic emphasis, or perceptual "loudness" (Wagner et al., 2015b). These concepts share the description of a perceptual impression pertaining to linguistic units (words, syllables, utterances), i.e., these units are perceived to "stand out" relative to their context (Terken, 1991). Grammatically, this impression of "standing out" can be linked to the prosodic organization (i.e., stress) acting within a particular phonological domain (e.g. phonological words, phrases), thereby forming "heads of prosodic feet", "lexical stresses", "nuclear accents", or a certain height in a metrical grid. While the impression of prominence is strongly intertwined with grammatically licensed prominence, these can belong to very different levels of grammar (phonology, morphology, syntax, semantics, pragmatics, discourse organization), and may vary strongly in nature (cf. section 1.2). Also, certain speech registers (such as shouted speech or Lombard speech) may be structural cues that have an effect on the phonetic features related to prosodic prominence. In such cases, a preference for a particular prominence pattern is caused by the situative or contextual embedding of on utterance, and may be at odds with the prominence patterns licensed by other levels of grammatical organization (Mooshammer, 2010). Therefore, the various signal and structural cues to prominence do not always go hand in hand, and listeners and speakers somehow need to manage the interface between these types of cues.

1.1. Signal correlates of prosodic prominence

The phonetic realization of prominence is language dependent (Andreeva et al., 2012; Barry et al., 2007; Rosenberg et al., 2012), may differ between dialects (Smith and Rathcke, in press, this Special Issue), and has been found to be related to pitch movement, height, and shape (Baumann and Röhr, 2015; Heuft, 1999; Mahrt et al., 2012; Niebuhr, 2008; Terken, 1991), duration and intensity (Fry, 1958; Kochanski et al., 2005; Turk and Sawusch, 1996), spectral emphasis and articulatory effort (Campbell and Beckman, 1997; de Jong, 1995; Heuft, 1999; Mooshammer, 2010), and prosodic context, e.g., the position of a pitch accent in the F0 contour (Gussenhoven and

Rietveld, 1988), or the prominence of adjacent syllables (Arnold et al., 2013). For Irish, Ní Casaide et al. (2013) found a compensatory trade-off relation between pitch features and voice source correlates of prominence: words were still perceived as accented even if they lacked clear-cut pitch excursions. In these contexts, voice source features related to prominence were found to be comparatively stronger. However, the exact interplay of signal cues (additive, compensatory) in the expression of prominence is still far from clear.

More recently, a number of investigations have pointed to a strong coupling between prosodic prominence and speech-accompanying gestures such as manual gestures or head movements (cf. Wagner et al. (2014) for an overview). There is mounting evidence for a strong parallelism in the production of acoustic prominence and simultaneous beat and deictic gestures (Krivokapic et al., 2015; Krivokapić et al., 2017; Leonard and Cummins, 2010; Loehr, 2012; Mendoza-Denton and Jannedy, 2011) which develops early in language-acquisition (Esteve-Gibert and Prieto, 2014). Convincing evidence for the existence of the cross-modal motor co-ordination on the level of prosodic prominence has also been put forward by Parrell et al. (2014): Using a co-speech tapping task, they show that speech-gesture coupling is not constrained to temporal alignment or movement duration, but that verbal emphasis even influences the magnitude of the corresponding manual movement. The temporal coupling between speech and gesture in prominence production is furthermore modulated by grammatically licensed prominence, i.e., the synchrony between speech and co-speech movements increases when new, unpredictable information is being uttered (Wagner and Bryhadyr, 2017).

1.2. Structural correlates of prosodic prominence

In addition to the obvious phonological categories such as lexical or phrasal stress, a wide range of structural linguistic features have been identified that may cue prosodic prominence, among these being phrasal position (Fougeron and Keating, 1997; Vainio and Järvikivi, 2006), information structure (Xu, 1999; Féry and Kügler, 2008), coreference or givenness (Baumann and Riester, 2013), informativeness (Calhoun, 2010), relevance and predictability (Aylett and Turk, 2004; Watson et al., 2008), lexical class (Widera et al., 1997), and lexical and syllable frequency (Bell et al., 2009; Samlowski, 2016). Not only do listeners expect individual linguistic items to be prominent because of their semantic or pragmatic function, they also show a tendency to expect (and perceive) prominence patterns as rhythmic alternations (Dilley and McAuley, 2008; Niebuhr, 2009; Vogel et al., 2015). Similar

findings have long ago been described for series of acoustically identical non-linguistic signals (isochronous pulse trains) (Bolton, 1894), and they go hand in hand with the postulation of traditional phonological constraints avoiding "stress clashes" or longer sequences without prominent units ("lapses") (Liberman and Prince, 1977). Clearly, many of the prominence predicting factors named above may conspire and are difficult to disentangle in running speech: New, relevant information tends to be less predictable, may be placed in syntactic positions prone to attract prominence (e.g. by topicalization) and is typically expressed with the help of intrinsically more prominent open class words. Besides, the named factors are subject to general prosodic constraints of stress and accent placement, e.g. the placement of prosodic focus may depend on other demands of the phonological grammar such as the adequate placement of (nuclear) accents or tonal realization (Calhoun, 2010; Turco et al., 2013; Xu, 1999). These complex interactions may also have an impact on how structural prominence is phonetically expressed (cf. Section 1.1). We have plenty of evidence that some structural prominences — e.g. focal accents — trigger a range of phonetic prominence cues, while other types of phonological prominence do not: For languages with fixed lexical stress, its phonetic expression may be weak at best, unless it coincides with sentence-level prominence (cf. Szalontai et al. (2016) for Hungarian; Ćwiek and Wagner (2018); Malisz and Wagner (2012) for Polish). This evidence hints to the possibility that linguistic units that are already made prominent by their structural location (e.g. topicalized, phrase-final, in stressed position) may be in less need of signally cued prominence as compared to prominences occurring in atypical positions, such as on function words. This assumption receives further support from evidence that narrow focus tends to be produced with more prosodic effort than wide focus (Baumann et al., 2006; Hanssen et al., 2008), and with investigations based on computational modeling (Kakouros and Räsänen, 2016), who found that prominence perception co-incides with low predictability regions within the speech signal.

1.3. Perceptual integration of structural and signal correlates to prosodic prominence

Given the many sources of influence and their complex interaction, prominence perception is often explained by listeners' ability to integrate their (linguistically guided) expectations, or top-down knowledge about prominence functions, with their auditory impressions which transport prominence-lending cues via the acoustic signal. The nature of this integration process

is hitherto little understood, but the impact of top-down expectations is apparently strong (Cole et al., 2010; Eriksson et al., 2001) and may depend on language experience, as it tends to have less influence when listening to non-native speech signals (Wagner, 2005; Eriksson et al., 2001). However, the amount of top-down expectations are variable and their impact may be a matter of a conscious choice (Cole et al., 2014). For the free word order language Russian, Luchkina and Cole (2014) report that both word order and signal cues independently mediate the impression of perceptual prominence. However, Turnbull et al. (2017) found that in American English, contrastive contexts increased perceptual prominence only if a highly prominent pitch accent type ($L + H^*$) was present. That is, discourse context alone cannot necessarily yield perceptual prominence independently of its phonological expression. Bishop (2016) describes that both signal and information structural cues have an independent effect on prominence perception, and information structural expectations of focus did not make listeners more sensitive to signal cues. However, the impact of structure apparently depended on listeners' pragmatic skills. In a recent study investigating the processing of prosodic prominence in German based on a large set of various types of potential predictors, Baumann and Winter (2018) identified pitch accent related cues (presence and shape) as having the highest impact on naïve listeners' prominence judgments. However, they also isolated two different listener strategies among their annotators: one group predominantly reacted to signal cues when annotating prosodic prominence, while another group was driven more strongly by structural cues. Bishop, Kuo, and Kim (in press, this Special Issue) confirm that cue integration is - among other factors - a matter of individual processing strategies. In a cross-linguistic study of cue integration, Cole et al. (2019) find language-specific strategies, with listeners paying most attention to those cues with the highest functional load in the respective language.

Despite these recent findings, we still know very little about the integration process of signal and structural cues, what the individual listener strategies within and perhaps across linguistic communities may look like, or how they evolve or adapt to situational needs. Given the complexity of the concept of prominence as it relates to structure, meaning, and cues from the speech signal, and given the fact that individual listeners may have their individual strategies of dealing with this richness (Baumann and Winter, 2018; Watson, 2010), it is therefore a necessity to study naïve listeners' impressions on a larger scale. Unlike expert annotators, they are unbiased by theoretical

preconceptions (Cole and Shattuck-Hufnagel, 2016) and are our only chance to get a more complete picture of the various potential language-specific and listener-specific strategies. Furthermore, our annotation protocols need to allow for rich individual variation (Cangemi and Grice, 2016). However, if annotations are to be gathered realistically on a larger scale, they will have to be accessible using methods that do not require a long training phase and are neither too cumbersome nor time-consuming.

1.4. Annotating prosodic prominence

Probably due to its different conceptualizations, there exists no standard or consensus approach for the annotation of prominence. Instead, a variety of annotation protocols have been proposed or used in the past, differing in (i) the level of annotation or prominence domain, (ii) the scale used for prominence annotation, and (iii) the way of how prominence judgments are averaged and normalized across several listeners (Wagner et al., 2015a). To this day, the majority of prominence studies rely on binary or unary impressions of prominence (Heckmann, 2014; Kalinli and Narayanan, 2009; Wang and Narayanan, 2007; Rosenberg et al., 2012), where units are divided into non-prominent and prominent ones based on the presence or absence of a perceptual impression or a strong signal cue such as a pitch accent. While these binary operationalizations have shown to be successful in many technical applications such as the automatic detection of relevant linguistic units, they forbid investigations that operate on different levels of prominence, e.g. differences between lexical stress and sentence accent, word-internal prominence relations (primary vs. secondary stress, compound stress)[1] or gradient prominence relations between accents. To achieve a more fine-grained assessment of prominence relations, Fant and Kruckenberg (1989) suggest a multilevel, quasi-continuous scale of syllabic prominence. In a related approach, Eriksson et al. (2001) introduce a continuous scale for prominence ratings, using GUI-based sliders to assess the prominence impressions for individual syllables. Other researchers employed scales with 11 (Malisz et al., 2015), 4 (Kügler et al., 2015), or 3 (Lacheret et al., 2013) levels of prominence. An alternative approach (Cole et al., 2010; Wightman, 1993) operationalizes continuous prominence annotations as unary impressions of prominence cumulated across several listeners, thus reflecting the probability of a linguistic

[1] Cf. Roettger and Gordon (2017) for a meta study on the lack of disentangling lexical stress and sentence accent in empirical studies of lexical stress.

unit to be perceived as prominent within a larger community (= "p-score"). With respect to the linguistic domain of prominence, most studies either measure prominence on the level of the syllable or word. Rather than predefining such a discrete level of annotation, Kunter and Plag (2007) used a slider-based approach to let listeners determine the exact location of prominence within English compounds. For an illustration of the most popular approaches to prominence annotation, cf. Figure 1.

Despite this striking heterogeneity, comparatively few studies have been specifically dedicated on the evaluation of these competing approaches. Portele et al. (2000) report on a high agreement between expert annotators trained on an annotation approach using 32 levels of prominence (Spearman-ρ between 0.7 and 0.8). Lacheret et al. (2013) find a good agreement for expert annotators using 3 levels of prominence, while Jensen and Tøndering (2005) argue that for word-based prominence distinctions, cumulating unary prominence impressions across several naïve listeners will reach similar results as more fine-grained expert annotations. Arnold et al. (2011b) systematically compared the efficiency of various multi-level prominence annotation schemata (4-, 7-, 11- or 31-level, and continuous scale), gathering prominence annotations using a slider-based GUI-approach. They concluded that multi-level schemata reflect richer impressionistic details (e.g. caused by contextual priming) and are not considerably more time-consuming than approaches with slightly fewer annotation levels. In their crowdsourcing study, Malisz et al. (2015) found a very high variation of prominence judgments across naïve annotators for a multi-level syllable prominence scale. Their annotators also reported the task to be difficult and cumbersome, which is in line with the conclusions by Jensen and Tøndering (2005) in favor of a simpler approach for naïve listeners.

Surprisingly little attention has been paid to the fact that all established annotation protocols are both listening and reading tasks, i.e., a stimulus is read *and* listened to, and prominence impressions are then assigned to the orthographic material. It has long been known that read stimuli prime expectations for speech that is listened to (Tanenhaus et al., 1980), and can even influence its phonological categorization (Escudero et al., 2008). It can therefore be expected that the bimodal stimulus presentation may have an impact on the annotation results, and even to some degree explain the strong impact of structurally cued expectations on prominence perception. It should be noted, though, that Cole et al. (2010) implemented some countermeasures to orthographically cued prominence, e.g. by deleting capitalizations and

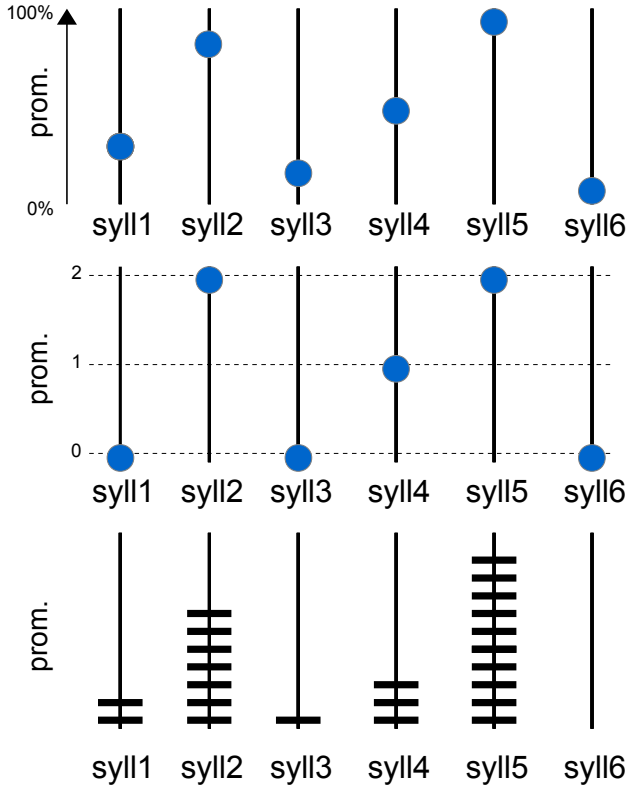

Figure 1: Schematic overview of three popular prominence annotation methods (from top to bottom): (1) Fine-grained continuous prominence annotation. (2) Less fine-grained prominence annotation using a discrete (here: three) amount of distinct levels. (3) Prominence annotation based on cumulative unary impressions across several annotators, where each bar represents an annotator's prominence marking.

punctuation marks in their orthographic transcriptions.

Lastly, so far very few studies exist that examine the influence of the prominence domain on the annotations themselves. Arnold et al. (2011a) showed that word-level prominence judgments are more reliable and in higher agreement with established acoustic prominence cues than syllable-based ones, indicating that word-level prominence judgments may be comparatively easier, or at least more reliable to annotate. Apart from this study, we have little evidence as to what role the level of annotation plays in the assessment and perception of prosodic prominence.

1.5. Speech-gesture co-ordination as prominence indicator?

It has been found that the strong speech-gesture link described in section 1.1 extends beyond production, as co-speech gestures have been found to independently contribute to the impression of prominence (House et al., 2001; Swerts and Krahmer, 2008) and to increase the intelligibility of speech (Al Moubayed et al., 2009). Gestures, or rather co-speech movements[2] that are co-ordinated with a signal can also facilitate the processing of musical rhythms (Phillips-Silver and Trainor, 2005, 2007), challenging verbal tasks such as reading comprehension (Llanes-Coromina et al., 2018b) or the production of acoustic-prosodic patterns in an L2 (Llanes-Coromina et al., 2018a). These findings highlight the strong cross-modality that accompanies the general processing of acoustic signals, both in perception and production, and challenge the seemingly clear-cut differentiation between structural and signal cues to prominence, as (motor) signal cues as such may actually trigger or strengthen structural expectations and vice versa.

A large body of interdisciplinary research furthermore describes that humans spontaneously and unintentionally synchronize or "entrain" their movement patterns to their interlocutors' verbal and non-verbal behaviors (Condon, 1974; Erickson and Shultz, 1982; Loehr, 2004; Richardson et al., 2007; Yun et al., 2012). While similar co-ordinative patterns are described among different kinds of animals, humans are the only species that has been found to synchronize these across various tempi, or even when lacking a regular

[2]Not all co-speech movements constitute communicative gestures in that they fulfill a particular communicative function. However, we assume that most of them may at least to some degree be typical *beat gestures*, i.e. contributing to, or encoding prosodic prominence. In this paper, we refer to all types of co-verbal movements as "gestural", without making any hypothesis as to their specific communicative function.

external stimulus such as a metronome or music (cf. Cummins (2011) for an overview). Building on these insights, Wagner et al. (2013) postulated their theoretical framework of Interaction Phonology, which regards a listener's motor co-ordination with the verbalizations of an interlocutor as a logistic component, or as a "rhythmic scaffold" that facilitates the perceptual processing of phonological units, boosts the comprehension of higher-level linguistic information and aids the management of the ongoing discourse.

While these theoretical assumptions are notoriously difficult to test (Beier and Ferreira, 2018), we still take them as support for the hypothesis that perceived prosodic prominence may at least to some degree be encoded in a kind of 'motor shadowing task'. We consequently expect that the amplitudes, durations or velocities of co-speech movements accessed in a listening task may provide a suitable indicator of perceptual prosodic prominence. We will further evaluate this hypothesis in the remainder of the paper.

1.6. Research goals and structure of this paper

As explained above, although a plethora of signal and structural cues to prominence have so far been validated, we still know very little about their integration in individual listeners. In order to shed light on this issue, prosodic research is in need of accessing naïve listeners' individual impressions of prosodic prominence. As prominence impressions and prominence relationships between linguistic units may be of a fine-grained nature, our annotations should be likewise. So far, no such annotation protocol exists. Our research goals for this paper are hence twofold: First, we will explore the suitability of a novel approach to prominence annotation that exploits the strong link between prosody and gesture in production, generates prominence annotations that are fine-grained, works independently of simultaneous orthographic input, and can be produced by naïve annotators. If the usefulness of such an annotation protocol could indeed be shown, we furthermore would have produced evidence supporting the hypothesis illustrated in section 1.5, namely that the process of speech perception includes some kind of internal "motor shadowing" or "motor co-ordination", which can be made use of in a prominence annotation task. We will furthermore explore, whether the domain of prominence annotation, namely syllable-level and word-level prominence, has an impact on the suitability of the annotation scheme. Second, we want to find out whether prominence annotations shed light on strategies of integrating various cues to prosodic prominence, and whether these differ across different annotation protocols and across different groups of listeners.

We will first address the possibility of a novel annotation scheme in a feasibility study. Next, we compare these gesture-based annotations with more established annotation protocols, namely the ones devised by Cole et al. (2010) and by Fant and Kruckenberg (1989), but also with a pure orthography-based "annotation" of linguistic expectations. Subsequently, we explore the weightings of signal-based and structure-based prominence cues across the different annotation protocols using Random Forest Models trained on average annotators, and lastly, we investigate, whether our gesture-based prominences reveal unique strategies for individual groups of listeners, followed by a general discussion.

2. Feasibility study: Can prominence impressions be accessed in a "drumming task"?

In order to find out whether it is feasible to exploit gestural movements to access prominence impressions, a combined perception and (drumming) production study was carried out. In this study, participants listened to short, isolated sentences and were subsequently asked to "repeat" their impressions by drumming them on an electronic drum pad, i.e. the repetitions were transformed into a gestural modality. The underlying assumption is that the impact forces of the individual drum beats capture the prosodic prominence of the linguistic unit the drum beat corresponds with. Thus, the impact forces of the individual drum beats serve as operationalization of perceptual prominence. Interestingly, a very similar measure of "visible energy", i.e. the amount of movement present in a manual gesture, has been proposed for the operationalization of prominence expression in American Sign Language (Tkachmann, Hall, Fuhrman and Aonuki, in press, this Special Issue). As articulatory manifestations of prominence may relate more or less strongly to different levels of the prosodic hierarchy (Oh and Bird, in press, this Special Issue), we assume similar effects to be the case for our "drummed prominences". Hence, one part of the study examined listeners' ability to "drum" syllables, another part to "drum" words.

2.1. Methods

2.1.1. Annotation material

The audio material that participants had to "repeat" in the drumming task was taken from the Bonn Prosodic Database (Portele et al., 2000), which contains prosodically annotated audio recordings of sentences and stories

read by three professional speakers. Twenty sentences read by each of the three speakers (60 sentences in total, cf. Appendix A) were extracted from the database for the main annotation study, with ten additional sentences serving as training material for the participants. All sentences were rather short and with one exception consisted of one major intonation phrase only. The sentences belong to a set of standardized and phonologically balanced sentences, typically used for the purpose of audio quality assessment or synthesis evaluation (Sotschek, 1984). When choosing which sentences to use for the experiments, care was taken to ensure that the three realizations of each sentence differed somewhat from one another in terms of prosodic realization. That way, linguistic structure is held constant across a part of the annotation material, while acoustic cues to prominence are varied. As our sentence material is not embedded in any linguistic context, our material cannot be meaningfully interpreted with respect to the potential impact of the pragmatic and semantic factors on prosodic prominence. Thus, our analyses of the interaction between semantic-pragmatic factors and perceptual prominence are severely constrained by our choice of material.

2.1.2. Participants and instructions

Nineteen native German speakers without any reported hearing or motor skill problems took part in this study (13 female, age range 20–58). A few of our participants had a linguistic background, but only two of them had some training in prosodic annotation. Both of these were assigned to the word drumming task due to the random procedure. We did not preselect the participants based on their musical skills, but 2 participants reported actively playing an instrument or singing regularly. Two participants were active drummers (1 semi-professional rock music drummer, 1 hobby performer of Western African percussion music). Based on a random procedure, ten participants (including the two active drummers) were presented with the syllable drumming task while the remaining nine were instructed to "repeat" their auditory impressions by drumming once per word, and to modulate their drumming strength based on how they felt that a word/syllable "stood out". To avoid potential biases, any linguistic (e.g. stress, accent, important words) or signal related (e.g. loud, high, long) terminology was avoided during the explanation of the task.

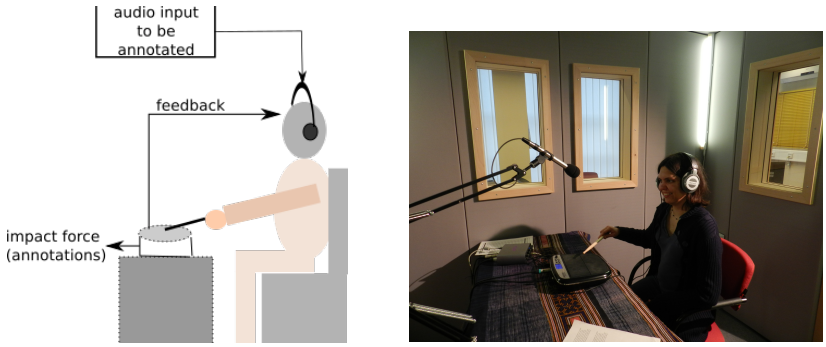

Figure 2: *Schematic overview of gesture-based annotation setup (left) and picture of an actual annotation session (right).*

2.1.3. Procedure

The annotations were performed with a standard electronic drum pad (Alesis SamplePad) in a sound-treated studio at Bielefeld University. The sensitivity for impact force was adjusted across all areas of the drum pad, and we chose a sample with minimal echo. Informal tests showed that impact forces are linearly transformed into the sample intensities (dB) of the output sample. Participants were presented first with the ten training sentences and afterwards with the 60 sentence recordings of the main study. The order of the training and test sentences was quasi-randomized for each participant, taking care that repetitions of the same sentence by different speakers were maximally far apart from each other. The participants were instructed to listen to each sentence over headphones and subsequently beat on the electronic drum pad once per perceived syllable (syllable task) or once per perceived word (word task), using a standard rock music drum stick (Maple, 5B). The drumming was performed *after* having listened to the audio presentation, as a real shadowing task where listeners drum while they are listening to an audio signal had shown to be too difficult in informal pilot studies. They were allowed to listen to sentence recordings again and/or repeat their drumming in case they were not satisfied with their performance. Participants were asked not to say the sentences aloud during the drumming, in order to have them concentrate on their perceptual impression. While drumming, participants monitored their performance via headphones (cf. Figure 2). Audio and MIDI output of the drum pad were recorded as well as the sentence stimuli which were played to the participants. The boundaries of the drummed sentence repetitions were annotated manually using Praat (Boersma and Weenink,

2019). By then automatically extracting the information encoded in the MIDI output (Walker, 2008) and comparing the MIDI time stamps with the relevant drum sounds in the audio file, it was possible to determine the impact force information stored in the MIDI file for each of the drum beats[3], i.e., an internal number on which the intensity of the output sound is based. That way, we could also determine the number of drum beats, which ought to be identical to the number of syllables or words in the corresponding sentence. Due to our sequential task (listening is followed by drumming), we lack a clear temporal alignment between the drumming patterns and audio signal. We are thus unable to relate the drum beats to corresponding syllables or words unless they are identical in number, and currently discard material where this is not the case.

2.2. Results

2.2.1. Coverage and time consumption

In order to estimate the difficulty and cognitive load of the drumming task, we analyzed the performances of five word drummers and five syllable drummers. As drum beats had to be attributed to the individual syllables or words, it was only possible to interpret participants' responses if the number of drum beats matched the number of words or syllables in the sentence. We use these mismatches as a first rough indicator of the task complexity. For the syllable drumming task, 16 of a total of 300 items (5%) had to be left out of the further analysis for this reason. The word drumming task resulted in a smaller meaningful output. Here, 38 of 300 items (13%) could not be interpreted. Thus, word level annotations appeared to be considerably more difficult to perform using the drumming task.

For both conditions, participants were able to go through the drumming task at a fast pace and only occasionally demanded to listen to or drum a particular sentence more than once. Including the time used for the training sentences as well as repetitions or sentences which had to be discarded from the analysis, the average time consumption per annotated sentence varied between 1.8 and 3.6 seconds for the syllable drumming task, and between 2.5 and 4.6 seconds for the word drumming task. This indicates a somewhat higher cognitive load for word drumming, as the sentences were identical

[3]Within the MIDI format, this impact force is called "velocity". We refrain from using this term due to its ambiguity.

across tasks, and fewer drum beats had to be produced in the word drumming task. The two participants with previous drumming experience were not the cause for the higher pace detected for the syllable drumming task.

2.2.2. Drumming patterns across participants and linguistic units

When investigating the individual drumming patterns both for words and syllables, we see a high inter-individual variation. In a first visual impression (cf. Figure 3), many of the participants appear to match their impact forces with well-known features of German utterance prosody, e.g. nuclear accents predominantly being appearing on nouns towards the right boundary of the prosodic phrase (the word "Süßigkeiten" in the example sentence), coinciding with lexical stresses (the syllables /zy:/ and /kaU/), with verbs and function words receiving less prominence than nouns, and with some drummers employing some kind of rhythmical alternation. Besides, the figures also show that in some cases, the individual drummers strongly "disagree", and may be following alternative strategies: word drummer 5 shows a flat drumming pattern throughout, with the exception of the last word which carries a strong rising boundary tone, word drummer 2 shows almost a reversed pattern than the others, while syllable drummer 2 places the "lexical stress" of "Süßigkeiten" after the actual stress location, possibly aligning the impact force with the maximal and late peak excursion. To get a clearer picture of the amount of inter-individual variation in the distribution of impact forces, and as a first check of whether the participants impact forces relate to prosodic prominence at all, we measured the distributions of their impact forces and plotted them against (manually labeled) accented and unaccented syllables/words per participant (cf. Figure 4). These distributions show that the majority of the participants indeed differentiate accentuation by increasing their impact force stronger for accented items, but that they do so (a) to very different degrees, and (b) that at first glance, several drummers fail to make a clear-cut distinction between accented and unaccented items. The distributions also show - not surprisingly - that participants use very different ranges and means of impact forces. Syllable drummer 6 shows hardly any variation throughout her performance. Interestingly, she is the participant with a background in rock music drumming. We have the suspicion that she interpreted the task rather as a syllable counting task and aimed at maximal consistency throughout her performance. Also, it seemed that her impact forces regularly exceeded the pad's maximal sensitivity. We therefore decided to exclude syllable drummer 6 from the subsequent analyses.

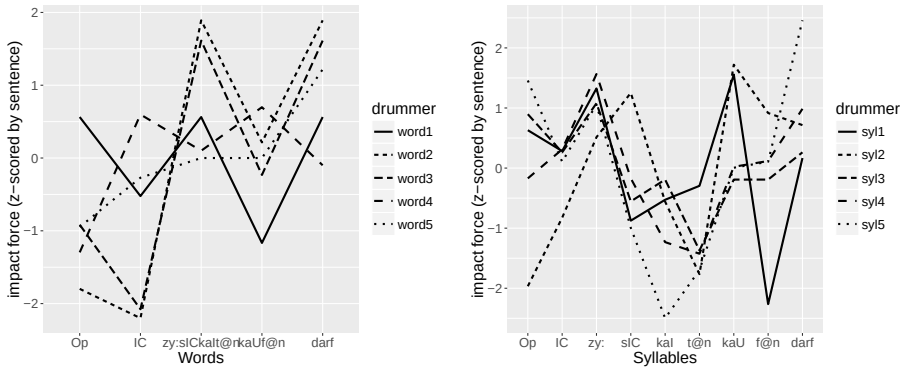

Figure 3: Z-scored (by sentence) drumming patterns for 5 word drummers (left) and 5 syllable drummers (right) for the identical utterance ("Ob ich Süßigkeiten kaufen darf?" – *Whether I sweets buy may?*).

As we are particularly interested in how our annotation protocol reveals the individual differences in interpreting prosodic prominence, we embrace the detected variations. Still, we need to ensure that the participants indeed followed some kind of a linguistic strategy. In oder to show the existence of non-randomness with respect to our linguistic items (words, syllables), we therefore calculated two linear mixed effects models using the R-package lme4 (Bates et al., 2015) both for our syllable and word drumming data. We entered the impact force as the dependent variable, individual items as fixed factors with 124 (syllables) and 89 (words) factor levels, and speaker and sentence (as well as word for the syllable model) as random factors. The resulting models (for the full models, cf. Appendix B) indicate that for syllable drumming, 59 out of 124 (48%) unique syllables and for word drumming, 33 out of 89 unique words (37%) make a clearly significant contribution on impact force ($|t| > 2$). Given the limited amount of annotated data per individual item, this result supports our assumption that the drumming task was carried out in a non-random fashion, as it is linked to the linguistic items that they were interpreting.

2.2.3. Inter-Drummer Agreement

In order to get an impression of the overall consistency among drummers, we calculated the intra-class correlations (two-way consistency ICC models on average units) for the word and syllable drummers. We closely followed the suggested procedure in Hallgren (2012) and used the R-package *irr* (Garner

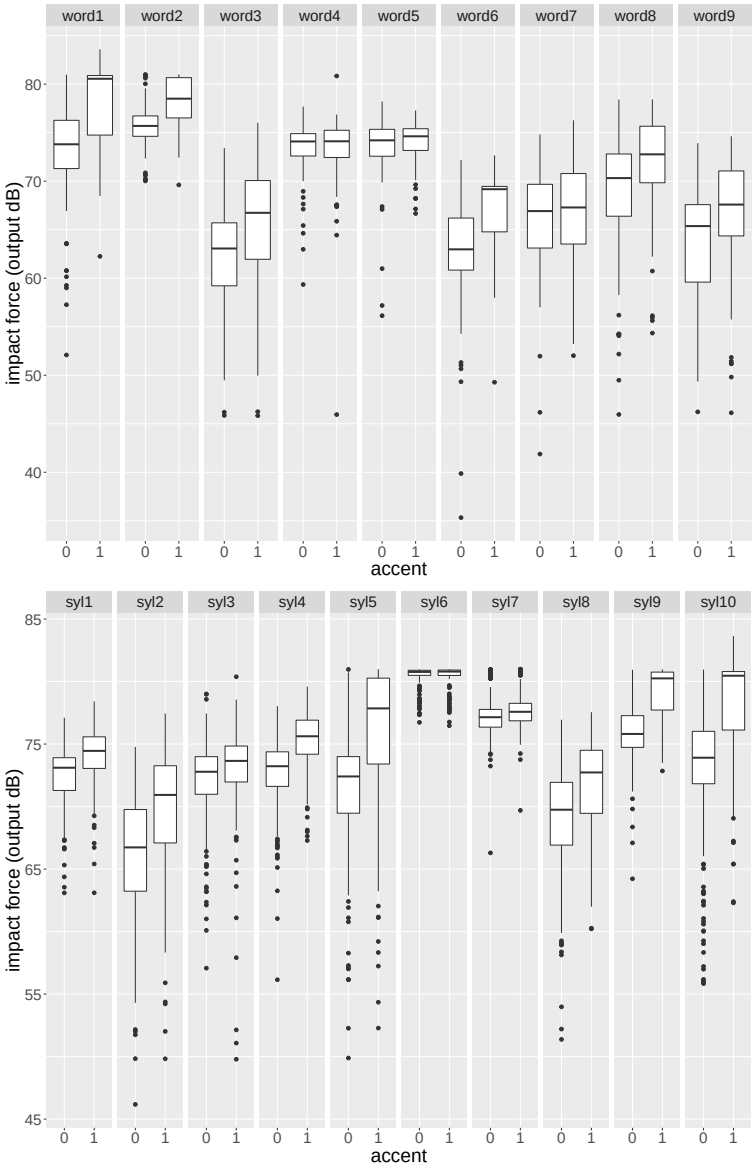

Figure 4: Distributions of impact forces for accented and unaccented words (top) and syllables (bottom) by participants (word1 to word9, syl1 to syl10). For a more intuitive interpretation, impact forces are given in the output dB of the drum pad.

et al., 2012). The resulting ICCs revealed high levels of consistency both for the syllable drumming task ($ICC = 0.72, F(389, 3112) = 3.63, p < 0.0001$) and the word drumming task ($ICC = 0.82, F(10, 80) = 5.44, p < 0.0001$).

We furthermore wanted to identify (dis)similarities between individual drummers' strategies for arriving at relative patterns of impact strength. To this end, we therefore calculated the pairwise-agreements between individual drummers on a sentence-by-sentence basis, using Spearman-ρ rank coefficients.

To prepare this analysis, all syllable- and word-level prominence ratings were z-normalized for each of the unique 60 sentences that had been annotated, to account for the different ranges in impact force. Next, Spearman-ρ correlation coefficients were calculated between the impact forces for each of the 60 sentence and each pair of drummers separately, resulting in up to 60 correlation coefficients per pair of drummers — due to the occasional performance "errors" (i.e. different numbers of drum beats and syllables), not all correlation analyses could be performed.

The medians of each of these pairwise correlations were entered in a correlation matrix, serving as a descriptive estimate of the pairwise agreement between individual word- and syllable drummers. The overall results are presented in Figures 5 (syllable drumming) and 6 (word drumming). All calculations were carried out within R, version 3.2.1 (R Core Team, 2015), visualizations were generated with the R package `corrplot` (Wei, 2013).

The analysis shows a wide range of similarities and dissimilarities between drummer performances for both tasks. For the syllable drumming, median correlations vary strongly (cf. Figure 5), ranging from high positive ($\rho = 0.79$) over practically no to even substantial negative ones ($\rho = -0.61$). Although word drumming appeared to be more difficult and error-prone than syllable drumming (cf. Section 2.2.1), the correlations between participants show a pattern similarly variable as the syllable drummers (cf. Figure 6), with some drummers agreeing strongly ($\rho = 0.83$), others poorly, and some even negatively ($\rho = -0.54$).

The lack of correspondence between many of our individual drummers indicates that individual participants paid attention to a different set of signal or structural cues when interpreting how strongly individual syllables stood out in the sentence. However, given that most drummers show clear positive correspondences to several other drummers, the results also indicate that groups of drummers may indeed follow similar strategies of cue integration. These analyses will be taken up later in order to model different strategies

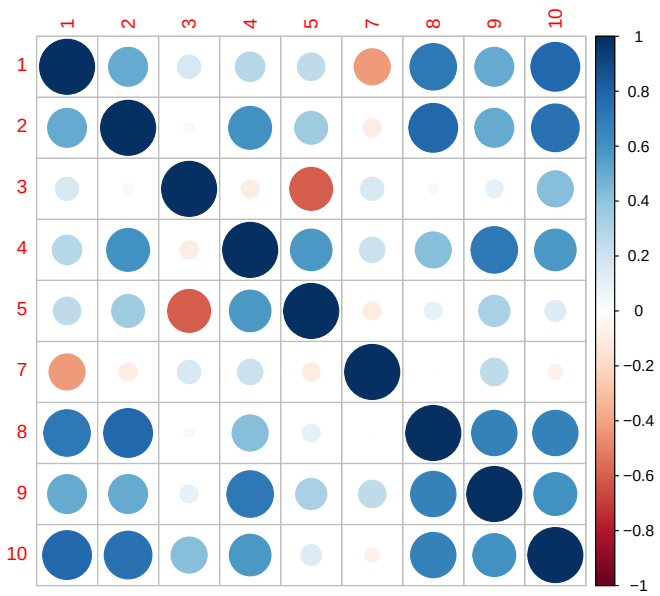

Figure 5: Medians of by-sentence Spearman-ρ correlations between sentence-wise patterns of drumming force across participants (syllable drumming task).

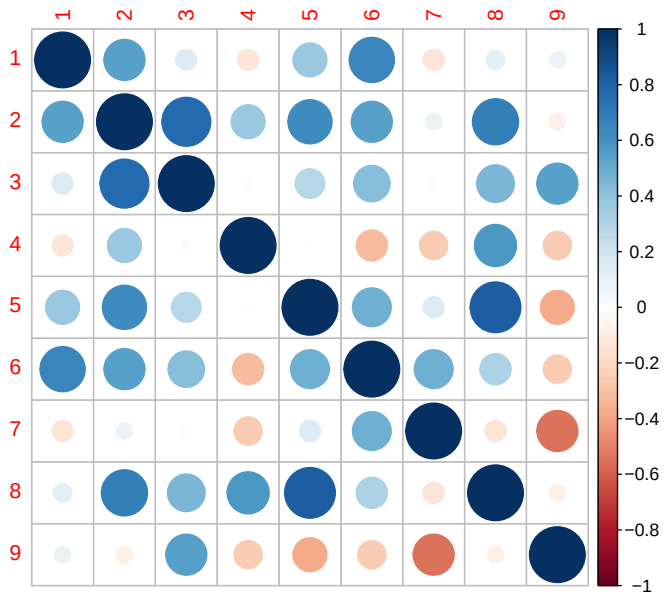

Figure 6: Medians of by-sentence Spearman-ρ correlations between sentence-wise patterns of drumming force across participants (word drumming task).

of cue integration (cf. Section 5).

2.3. Discussion

As hypothesized, we found that the drumming task allows for a very fast and intuitive way to gather naïve listeners' impressions of previously heard utterances. In fact, the procedure allows for an annotation speed close to real time and with a training phase of a few minutes only. Due to this effectiveness, the proposed method appears to be suitable for the fine-grained manual prominence annotation of large corpora.

Due to the fact that drum beats and annotated units could only be aligned when they were identical in numbers, the annotation method still produces missing data, which could be overcome if the annotation was performed as a real-time shadowing task. However, this appears to be do difficult without considerable practice or alternative methods of assessing the movements. It thus seems promising to investigate more subtle measurement techniques or explore approaches for training annotators.

Interestingly, word drumming was considerably more error prone and time consuming than syllable drumming, leading to believe that the coordination of hand movements and words was comparatively more difficult and needed more cognitive resources. The reason for this may be that speakers may be predominantly used to a word-level prosody-gesture alignment only for highly prominent words carrying a pitch accent (cf. Section 1), but not for less prominent words. Another potential reason for the comparatively easier syllable-level temporal alignment between drumming and linguistic units may be that German syllable units show considerably less temporal variation than words, and are repeated within duration cycles roughly corresponding to participants' *eigenfrequencies* found in self-paced tapping experiments (cf. Repp (2005) for an overview). In fact, our approach to annotation can be regarded as a special kind of sensorimotor synchronization task. Thus, the syllable may simply provide a more suitable temporal scaffold for speech-motor coordination — a circumstance also made use of in German elementary school classrooms, where the concept of a syllable is often taught in clapping games. The better suitability for speech-motor coordination of syllables (rather than words) may be further emphasized by some features of the annotated language: German allows for complex compounding, which show a certain ambiguity with respect to the number of words they actually represent. Another result of compositional, derivational and inflectional morphology, words are of an indefinite phonological length. We thus cannot

rule out that the higher cognitive load found for word-level annotations is a language specific effect.

Typically, the usability of standardized prosodic annotation protocol is evaluated based on inter-annotator agreement (e.g. Pitrelli et al. (1994); Kügler et al. (2015)). Our ICC-based analysis confirms the suitability of our proposed scheme, when averaging across various annotators. However, the correlation analyses revealed a high degree of inter-individual differences among our drummers. The finding that some participants seem to follow altogether different strategies could be used as a criticism of the proposed method. However, since our method does not define a standard approach for a prosodic expert annotation with a set of well-defined annotation criteria and available training materials, we do not consider this a major drawback. Rather, our approach welcomes the inter-individual variation, as we are interested in how individual listeners pay attention to various aspects of phonetic detail or linguistic structure, all of which may contribute to the overall impression of prominence.

Moreover, our models showed a clear impact of many individual items (syllables or words) on impact force, thus confirming that at least, our participants did follow strategies that were connected to a linguistic assessment of the acoustic data they were interpreting. Furthermore, an analysis of non-expert impressions may substantially contribute to our knowledge of the signal-structure integration process, as we so far typically have to rely on expert annotations whose judgments are to some extent driven by theoretical (rather than impressionistic) considerations, thus being endangered by a certain potential for circular reasoning.[4]

One remaining caveat is that the sentences annotated with the novel method consisted of carefully read, short sentences typically comprising a single intonation phrase. It remains unclear whether our method is straightforwardly applicable to more spontaneous data, consisting of longer and more complex intonation phrases, and containing typical phenomena of spontaneous speech such as self-corrections or hesitations. A second remaining caveat is that we did not control for our annotators' ability in performing the task. It is possible that some motor-coordination talent is needed to

[4]Obviously, certain analyses call for expert judgments, e.g. the qualitative assessment of a particular accent type needs insights into phonological concepts and distributional criteria.

conduct the drumming exercise. It is further conceivable that the few annotators who hardly correspond to the others simply lack this talent. Follow-up studies need to thoroughly investigate, how drumming performances are actually influenced by skills, talents or experiences that are not related to the linguistic task of listening. However, we feel that the many advantages of this approach, e.g. the fact that no intensive training is needed, and that the annotation works without orthographic representations, outweigh these shortcomings.

3. Comparing gesture-based prominence annotations with established annotation methods

Apart from the issue of whether naïve annotators can manage a gesture-based prominence annotation (henceforth "drummed prominences"), we need to examine how well the obtained annotations correspond to more conventional prominence ratings, for which researchers have found a range of corresponding structural and signal cues. This is crucial in order to find out whether results stemming from more traditional prominence annotation methods are actually comparable to those resulting from our novel method, or whether they capture impressions of a different quality. Here, two popular alternative approaches for prominence annotation were compared to the drumming method: expert judgments using a fine-grained, quasi-continuous scale of prominence impression (henceforth "fine-grained expert"; cf. Fant and Kruckenberg (1989); Eriksson et al. (2001); Arnold et al. (2013)) and the method relying on cumulative unary impressions of prominence (henceforth "cumulative naïve"; cf. Cole et al. (2010); Wightman (1993)). As we are also interested in the integration process of structural and signal cues to prominence, we introduce a third "quasi-annotation", which is entirely based on orthographic material, and simulates the cumulative naïve method (henceforth "orthographic cumulative"). The material thus gathered is an approximation to "pure" structural top-down expectations, as participants provide their "quasi-annotations" without any access to signal cues (leaving aside the question of whether they generate an internal silent prosody).

Also, we will generally assess whether various established and novel approaches to prominence annotation indeed measure comparable impressions or not. This is crucial to re-assess a plethora of empirical work so far dedicated to the study of prosodic prominence, and can be considered an important methodological side aspect.

3.1. Methods

To enable a comparison of gesture-based and conventional annotation methods, we first computed a "median drumming profile", based on the median of all drumming intensities of the individual syllables or words, after these had been z-score normalized by sentence and speaker. For the novel method to be useful, we hypothesize that this "median drummer" yields a more or less representative picture of the prominence impressions within a linguistic community. Obviously, the collection of a larger sample and a variability analysis across different participant subsets of varying sizes would be necessary to verify this preliminary assumption (e.g. by following Cole et al., 2017).

For all 60 sentences annotated with the drumming task, we also obtained fine-grained expert annotations from the Bonn Prosodic Database (Portele et al., 2000). This database was already the source for obtaining the sentences for the drumming task and contains syllable-based prominence impressions from three prosodic experts for each sentence. The prominence annotations were carried out on a fine-grained scale ranging from 0 to 31, with annotation values > 20 roughly corresponding to highly prominent, often pitch accented syllables (Heuft, 1999). The expert annotations showed a high correlation, similar to the rest of the database (Spearman-$\rho > 0.7$), but care was taken to introduce some variability of prosodic production across sentences. We calculated the median prominence profiles of these three individual impressions to obtain median fine-grained expert impressions. As the prominence annotations in the database are based on syllables as the reference unit — for every syllable there exists one prominence impression reflected by a number between 0 and 31 — word prominences need to be derived in order to compare them with the word prominences obtained in the drumming task. This was achieved by using the maximally prominent syllable per word as an indicator of perceived word prominence. Even though this simplification may not be entirely correct (Arnold et al., 2012), we take it as a reasonable approximation of perceptual word prominence obtained with a fine-grained scale.

The same 60 sentences were further annotated using the cumulative naïve prominence annotation method, closely following Cole et al. (2010): The sentences were presented both orthographically and acoustically (via headphones) to 40 naïve annotators in a random order, who were then asked to underline those linguistic units they perceived as "standing out". All annotators performed the annotations alone, sitting in a quiet room. Twenty

annotators were asked to underline syllables (14 female, age range 19–37 years, median = 23 years), a further 20 annotators were asked to underline words (12 female, age range 20–41 years). All annotators were students who were compensated monetarily for their participation in the study. In the syllable annotation task, syllable boundaries were indicated within words using the symbol "–". For each sentence, the 40 individual prominence profiles thus gathered were then cumulated into fine-grained "p-scores" of prominence impression and normalized into a value between 0 and 1, reflecting the proportion of annotators who marked a word or a syllable as prominent.

Lastly, we collected orthography-based "prominence annotations", by presenting the 20 linguistically unique sentences in their unmodified orthographic form to 17 naïve "word-level annotators" (11 female, age range 19–26 years), and in a syllabified version identical in form to the one used for the cumulative impressions to 15 naïve "syllable annotators" (nine female, one unspecified, age range 19–28 years). Their instruction was to underline all syllables/words that "they would expect to be prosodically highlighted". All annotators were undergraduate students of linguistics with basic knowledge in phonetics and phonology, so they had a fundamental grasp of the concept of prosodic highlighting. However, none of them had any specialized prosodic training. They took part in the study voluntarily and were not compensated for their efforts. The annotations took place in a quiet classroom. These annotations were then cumulated into fine-grained patterns of prominence expectations and normalized into a value between 0 and 1, similar to a "p-score", reflecting the proportion of annotators who expected a word or a syllable to be prominent.

Subsequently, "drummed", "fine-grained expert" (experts), "cumulative naïve" (p-scores), and "orthographic cumulative" (orth) annotations were compared both on the syllable- and the word-level for mutual similarities (cf. Figure 7 for an example comparison based on one sentence). To this end, the annotation data were aggregated for word and syllable based annotations separately. Next, we calculated linear mixed effects models, with annotations gathered by one of the annotation protocols and annotation units serving as dependent variable, while the remaining annotation protocols were set as fixed factors (intercepts). The variables sentence, speaker and word (as well as syllable for syllable drumming) were entered into the models as random factors.

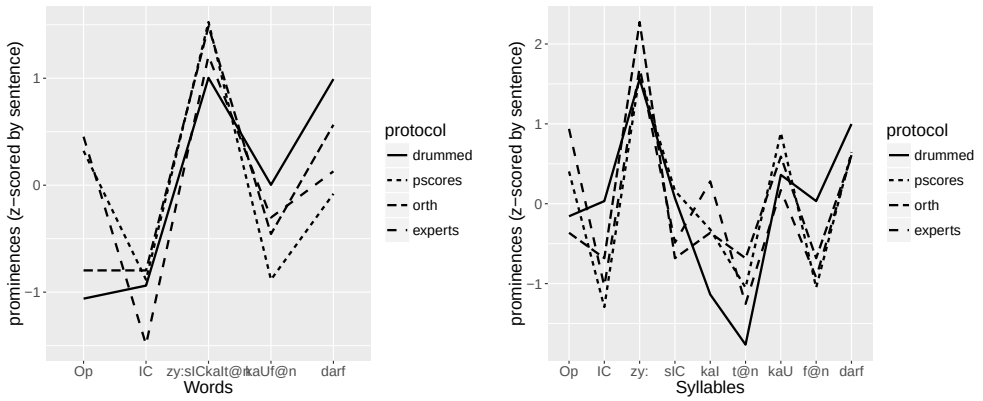

Figure 7: Example sentence with results for z-scored median prominence annotations gathered with various annotation protocols for words (left) and syllables (right).

dep. variable:	p-scores	orth	drum	experts
p-scores	-	**2.56**	**5.40**	**2.56**
orth	**2.9**	-	**−2.24**	**−8.03**
drum	**6.23**	**2.18**	-	**4.66**
experts	−0.11	**−5.98**	0.43	-

Table 1: Overview of linear mixed effects models (fixed factors) predicting word level prominence annotations (dep. variable, left column) by alternative annotation protocols (remaining columns). Cells show the t-values pertaining to the fixed factor, with $|t| > 2$ marked in boldface.

3.2. Results

As indicated in Table 1 and Figure 8 for word-based annotations, and Table 2 and Figure 9 for syllable-based annotations, the protocols stand in a strong linear correspondence to one another. The only pair of annotation methods failing to show such a correspondence are word-level expert prominences as predicted by p-scores and drummed prominences.

3.3. Discussion

The most important outcome of our comparison of novel and established methods for prominence annotation is that *all* procedures yield highly comparable results, i.e. each annotation protocol is predictive of the others. These are very good news to the prosody community, as they justify it to further compare empirical results based on different methods of prominence annotations. The only annotation scheme for this was not the case (expert

dep. variable	p-scores	orth	drum	experts
p-scores	-	**6.69**	**8.44**	**8.45**
orth	**10.66**	-	**4.17**	**10.66**
drum	**4.15**	**8.40**	-	**5.92**
experts	**7.53**	**10.54**	**6.09**	-

Table 2: Overview of linear mixed effects models (fixed factors) predicting syllable prominence level annotations (dep. variable, left column) by alternative annotation protocols (remaining columns). Cells show the t-values pertaining to the fixed factor, with $|t| > 2$ marked in boldface.

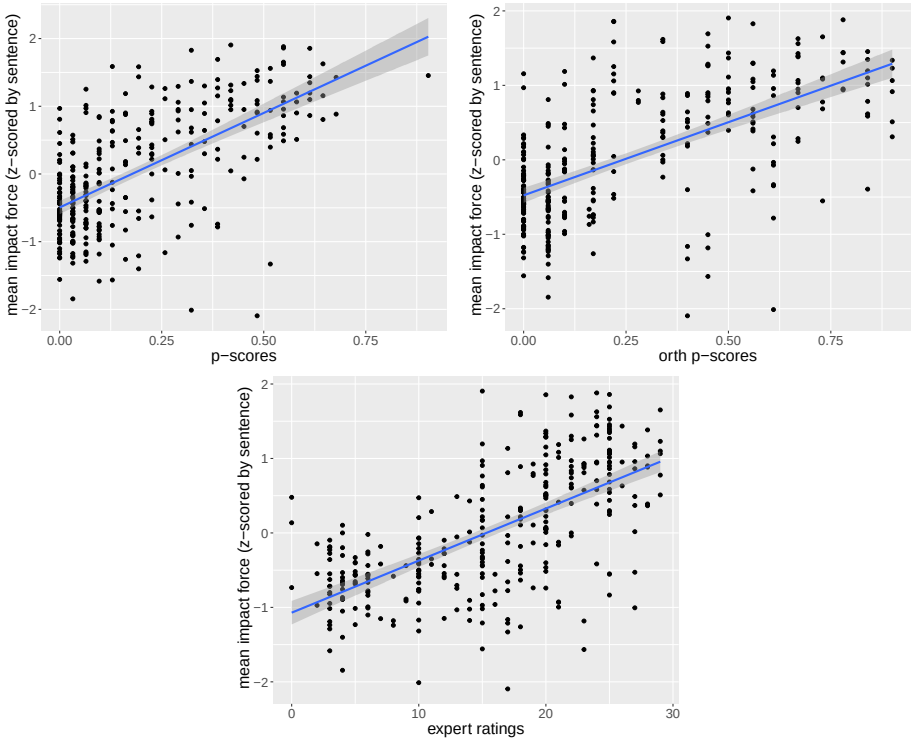

Figure 8: Scatterplots of impact forces gathered with the word-level drumming task in relation to p-scores (upper left), orthographic 'p-scores' (upper right) and fine grained expert annotations (lower).

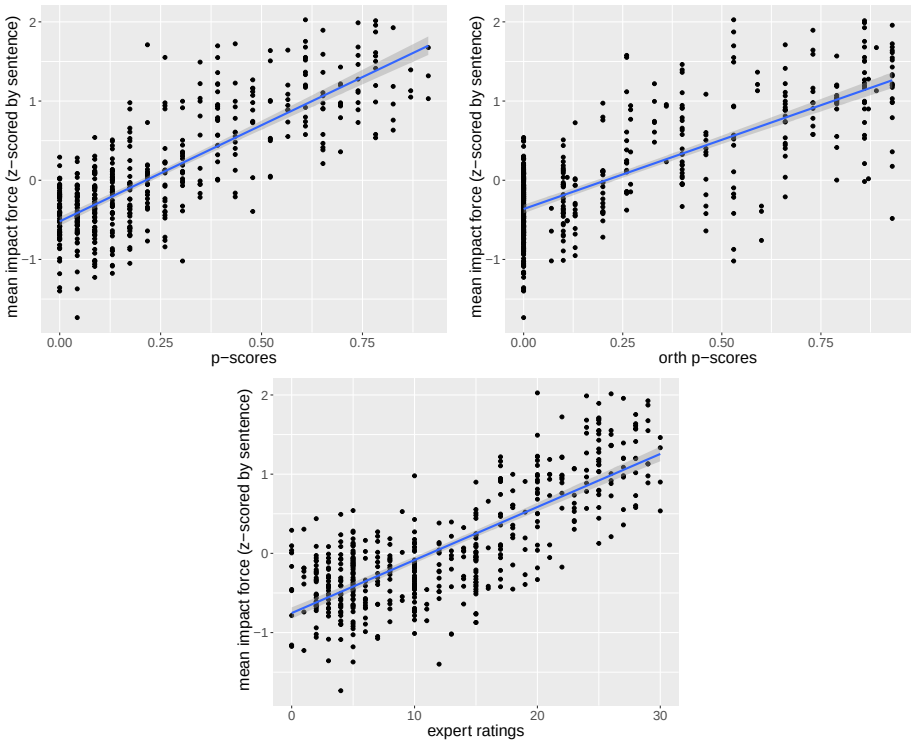

Figure 9: Scatterplots of impact forces gathered with the syllable-level drumming task in relation to p-scores (upper left), orthographic 'p-scores' (upper right) and fine grained expert annotations (lower).

"

word-level annotations), was derived from syllable-level annotations. We take this as evidence that our simplifying assumption, namely that the most prominent syllable within a word gives us a good approximation of word-level prominence, is inadequate. Obviously, prominence impressions cannot be straightforwardly compared across different levels of annotation.

Also, our study revealed that relying on naïve annotators is not a drawback at all, but is able to yield results highly comparable to those gathered by prosodic experts, as long as a substantial number of annotators (here $n = 9$) are available, across whom we can calculate average or cumulative impressions.

The "annotations" gathered from the orthographic material showed a strong correspondence to all other annotation protocols, thus confirming the often described strong impact of top-down expectations on prominence perception across methods. This persisted to be the case for the drumming task. Even lacking orthographic input during the annotations, it does not neutralize the impact of top-down expectations on prominence judgments. However, for word-level prominence annotations including orthography-based anntations, the negative t-values indicate that orthographic prominences are occasionally reversed, probably resulting from some divergence between prosodic realizations and prominence predictions based on text.

As our analyses have thus confirmed the general comparability of the drumming method to obtain meaningful prosodic annotations, we will subsequently evaluate the degree to which annotations gathered that way are indicative of an integration of structural and signal cues to prominence perception.

4. Investigating the integration of structural and signal cues to prominence

Since our previous analyses revealed potential differences in the integration of structural and signal cues in the different annotation procedures tested, we now further investigate whether such an assumption can be supported empirically. We will analyze these integrations separately for word-level and syllable-level annotations across the various protocols.

4.1. Methods

We train a set of Random Forest Regression Models (Breiman, 2001) based on our three different annotation protocols averaging over several an-

notators (median drummer, cumuluative p-scores, median expert) as dependent variables, and a set of well-established structural and signal correlates of prosodic prominence as predictor variables, both for word and syllable-level annotations. Beyond more traditional predictive models, Random Forests reveal a ranking of the importance of the various predictor variables, i.e., they deliver more information than mere model accuracy, the significance or effect size of a predictive factor. That way, they provide us with an exploratory tool for determining a hierarchy of the various structural and signal cues of prosodic prominence. Random Forest Models have further advantages over other types of predictive modeling, most notably their relative immunity against multicollinearity. This is of major importance since we can assume that many of our predictor variables are highly correlated due to the "conspiracy" and linking of a plethora of prominence related factors described in Section 1.

Before training our Random Forests, we chose a set of predictor variables. Perhaps somewhat unusually, we initially hesitated to include the predictor variable "presence/absence of a pitch accent", as it seems to be of an ambiguous nature: It may be regarded as a signal cue to prominence, as it manifests itself as prominence-lending pitch movement. Still, in many respects it is a structural phonological feature, heavily corresponding with structurally conditioned prominence such as phrasal position. We therefore felt that its usage as predictor variable would not be straightforwardly informative in determining the differentiated impact of structural or signal cues. However, as its impact on prominence perception (p-scores) has shown to be of major importance in recent work relying on a similar analysis of Random Forest Models (Baumann and Winter, 2018), we built a separate set of Random Forest Models including (1) the presence of a pitch accent (ACCENT) as a predictor, but left it out in a second stage of analyses. That way, we can determine whether the results by Baumann and Winter (2018) can be replicated on different data and with respect to various annotation protocols. Our pitch accent annotations were manual expert labels present in the original database.

As signal-based predictor variables, we selected well-established acoustic correlates of prominence, namely (2) syllable duration (SYLDUR), as well as (3) an integrative measure of acoustic prominence (ACPROM), that combines pitch movement, nucleus duration, intensity, and spectral balance into a single prominence value (Tamburini, 2006) and has been shown useful to model signal-level German prominence, e.g. for the purpose of model-

ing prosody in speech synthesis applications (Windmann et al., 2010; Malisz et al., 2017). It has been also established that its actual relationship with perceived prominence is imperfect, given that prominence is also driven by structural features (Tamburini and Wagner, 2007).

As paradigmatic structural variables, we chose (4) the phonological stressability of a syllable (STRESSABLE), basically distinguishing reduced syllables from non-reduced ones. This feature is only distinctive for syllable-level prominence and disregarded in the Random Forests predicting word-level prominence. Another paradigmatic structural variable is the (5) part-of-speech of the word (the syllable is contained in) (POS). Furthermore, we included a syntagmatic signal variable that may trigger the expectation of a prominence pattern due to rhythmic alternation, namely (6) the acoustic prominence of the previous syllable (PACPROM). For a detailed overview of the predictor variables, cf. Table 3. Obviously, many more potential predictor variables could have been chosen for a full exploration of prominence related cues, e.g. those related to the contextual predictability or information structure. Also, a study interested in whether listeners pay attention to different signal cues across varying contexts or listening situations, may find it useful to further decompose the signal cues to prominence. However, since our main intention is to get a global impression of whether structural or signal cues dominate the perception of prominence rather than building the best predictive model, and since our linguistic material is not well-suited for an analysis of more complex pragmatic or semantic influences anyway, we stick to this comparatively simple set of well-established predictor variables.

Predictive Random Forest Models were then trained on the 60 annotated sentences based on the 3 different annotation approaches and the 2 different levels of prominence annotation, with and without the predictor variable "ACCENT", yielding a total of 12 different models. Training was conducted using the `cforest` function contained in the `party` package (Strobl et al., 2007) within R (R Core Team, 2015). This procedure follows the suggestions by (Strobl et al., 2009), as the predictor variables were of different types and contained correlated predictor variables. Model accuracy was then assessed by obtaining the out-of-bag (OOB) predictions provided by the `party` package in R, and by then calculating a root mean squared error between OOB predictions and actual annotations.

predictor variable	**description**
ambiguous:	
ACCENT	presence of a pitch accent (manual expert annotation derived from the database Portele et al. (2000))
paradigmatic signal-based:	
SYLDUR	raw syllable duration (ms)
ACPROM	acoustic prominence, weighted sum of *nucleus duration*, *spectral emphasis* (Fant et al., 2000), *RMS intensity*, TILT-model based *pitch movements* (Taylor, 2000); cf. (Tamburini and Wagner, 2007) for details
syntagmatic signal-based:	
PACPROM	acoustic prominence (see above) of preceding syllable
paradigmatic structural:	
STRESSABLE	differentiates between phonologically reduced /ə,ɐ/ and full vowels
POS	differentiates between nouns (NOUN), adjectives (ADJ), full verbs (VRB), auxiliaries (AUX), adverbs/modal particles (ADV/MOD), prepositions (PREP), articles (ART)

Table 3: Overview of signal-based, paradigmatic structural and syntagmatic structural predictor variables used in the Random Forest Classification

4.2. Results

Generally, the models containing the predictor ACCENT perform very differently depending on whether they are applied on word-level (cf. Figure 10) or syllable-level (cf. Figure 11) prominences. For cumulative naïve annotations gathered as p-scores, the ACCENT predictor outperforms the others, thus replicating results by Baumann and Winter (2018). However, ACCENT ranks behind POS and SYLDUR for drummed prominences, and behind POS for expert annotations in their predictive power. For syllable-level annotations, however, ACCENT outranks the remaining predictor variables across all tested annotation protocols. It is noticeable, that the inclusion of ACCENT hardly improves the model accuracies, with the exception of fine-grained expert annotations on syllable-level.

When comparing the word-level with the syllable-level models, we see that the structural predictor POS ranks higher across all word-level annotation protocols, while signal-level predictors have a tendency to rank higher in syllable-level annotations. On the word-level, expert annotations are in fact best explained using the concept of POS, independently of the presence or absence of an accent.

There is also a tendency for drummed annotations to rank the predictor SYLDUR higher than the predictor ACPROM. This appears to be a systematic difference to the other annotation protocols, placing a stronger importance of ACPROM.

Across protocols and annotation units, the Random Forests were able to explain a substantial amount of variance present in the data, but are less accurate than Random Forest Models for prominence prediction that are reported in the literature (Arnold et al., 2013). However, these were trained on five times more data (i.e., the entire Bonn Prosodic Database described in Portele et al. (2000)), and on a larger set of predictor variables, including highly predictable contextual prominences.

4.3. Discussion

A comparison between the predictor rankings of the word- and syllable-level annotations reveals that POS is a more important cue for the perception of word-level prominences, while the prosodic cues appear to be more important for the perception of syllable-level prominences. This general tendency seems stable across annotation protocols and thus provides ample evidence that it indeed matters on which annotation level prominences are gathered.

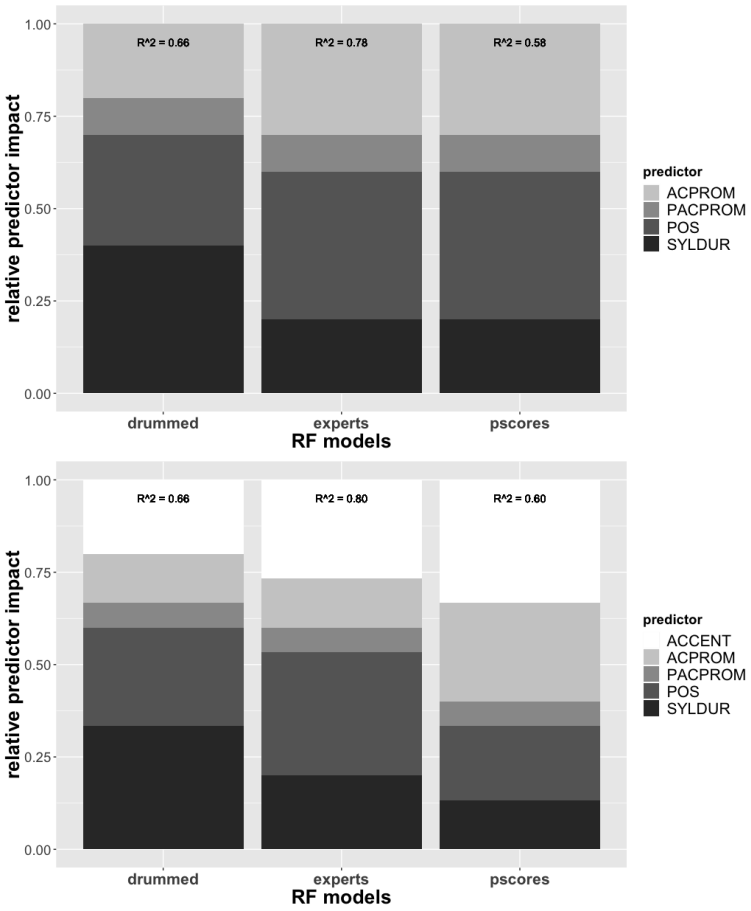

Figure 10: Importance ranking of predictor variables in Random Forest Models predicting word-level prominence annotation without (top) and with (bottom) the predictor ACCENT. For each model, predictive accuracies are given as RMSE (R^2)

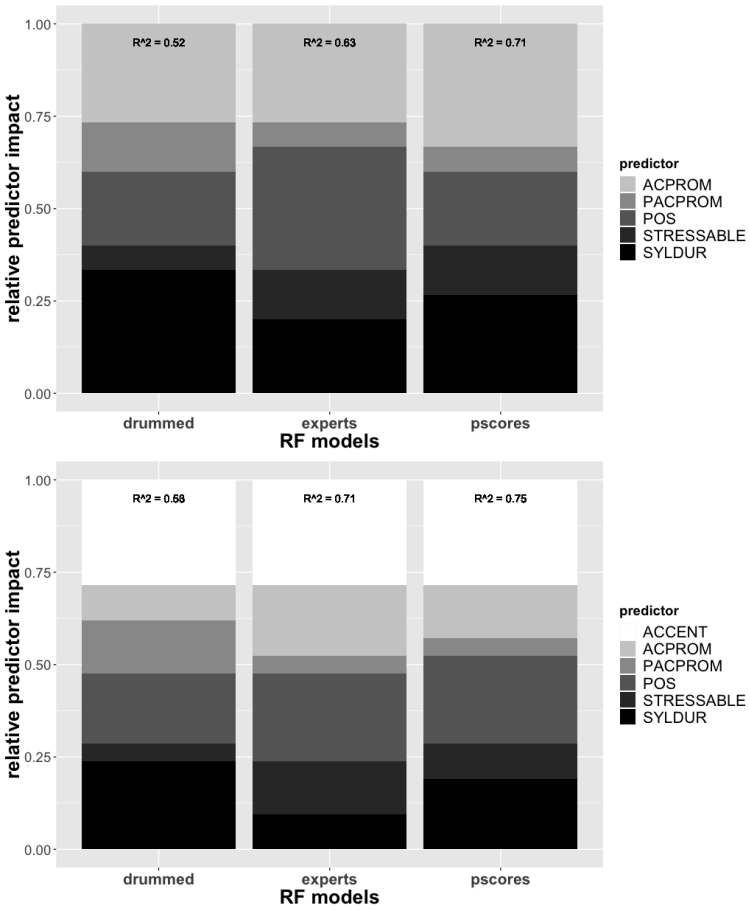

Figure 11: Importance ranking of predictor variables in Random Forest Models predicting syllable-level prominence annotation without (top) and with (bottom) the predictor ACCENT. For each model, predictive accuracies are given as RMSE (R^2)

When comparing the different annotation protocols, we furthermore see an influence on the method: The drumming task makes the participants rely less strongly on POS, which could be indicative of a less strong influence of orthographic input using this approach. Another difference of the drumming method appears to be that the annotators place comparatively more attention to syllable duration, it is possible that the influence of this cue is strengthened by the motor task, as the drum beats need to be interpreted in a sequential manner, placing an emphasis on the (syllable length) intervals between them. Interestingly, experts appear to be driven comparatively more by the cue POS than naïve annotators, as revealed across annotation levels.

When comparing the models with and without the predictor variable ACCENT, we see that except for expert syllable-level annotations, the absence of this predictor does not affect the model accuracies strongly, i.e. the annotators' behavior can be almost equally well predicted based on the remaining cues. From the strong positive impact of the predictor variable ACCENT on experts, we conclude that perhaps expectedly, experts use an abstract notion of accent in a much more systematic fashion than naïve listeners. Interestingly, ACCENT appears to be a much stronger predictor of syllable-level prominence, clearly dominating across all annotation protocols. This may be due to the fact that the POS information is less influential on the syllable level. On the word level, ACCENT remains to be the dominant predictor for p-scores, thus nicely replicating results by Baumann and Winter (2018) on a different data set, and with a different set of predictor variables. However, the different behaviors of word-level p-scores in comparison with the other annotation protocols are for the moment difficult to interpret, as the p-score models with and without accents show very different rankings (while drumming and experts stay similar in their rankings of the predictor variables).

Based on the model accuracies, we can also see that the drumming task introduces more noise than the cumulative naïve method and the expert annotations. It should be kept in mind, though, that there were considerably less drummers involved than annotators for the cumulative method. For now, we expect the drumming models to become more stable with a larger set of annotators. Model accuracies are more or less comparable for the annotations gained with the cumulative method relying on naïve annotators, and the fine-grained expert annotations, at least for the syllable-level annotations. This strengthens approaches relying on naïve annotators in prosody research,

provided the annotations are carried out by a sufficient number of listeners.

In sum, we see from our comparison that while producing similar results, all annotation protocols bring in their own idiosyncrasies, pointing to an effect of the annotation protocol on prominence cue integration. We also see that the level of annotation strongly influences the participants' behavior, so it is not advisable to extrapolate prominence impressions gathered on one level to those on another level.

5. Investigating individual listener strategies in integrating structural and signal cues to prominence

So far, our analyses were based on average or cumulative ratings of several annotators that we hope to approximate a strategy representative of a larger linguistic community. However, speech-based communication always takes place between individuals, who may follow their own individual strategies of cue integration or weighing. We will therefore try to understand better, whether we can indeed isolate different general strategies. For this endeavor, we again use Random Forest Classifiers trained on a set of established structural and signal cues known to be correlated with the perception of prosodic prominence (cf. Section 1). Rather than training one Random Forest on average impressions across all annotators, we will now apply them to groups of annotators whose annotation patterns are most similar. The Random Forests' importance ranking will then reveal an insight into individual integration processes. We will base the Random Forests on the gesture-based, syllable-level drummed annotations. Due to the fact that our word-level annotations do not deliver enough data points for calculating meaningful Random Forest Models, we restrict ourselves to syllable-level annotations.

5.1. Methods

Prior to training Random Forests, and similar to Baumann and Winter (2018), we performed a cluster analysis on the annotations from nine of our ten original "syllable drummers". This is based on the pairwise comparisons of syllable drummers shown in Figure 5. The medians of each of these pairwise comparisons were entered in a correlation matrix, which was transformed into a matrix of euclidean distances. This then served as input for the hierarchical clustering algorithm implemented in the R-package `cluster` (Maechler et al., 2015). The output of this process revealed two main clusters of annotators (cf. Figure 12), one cluster with six annotators (annotators 1,

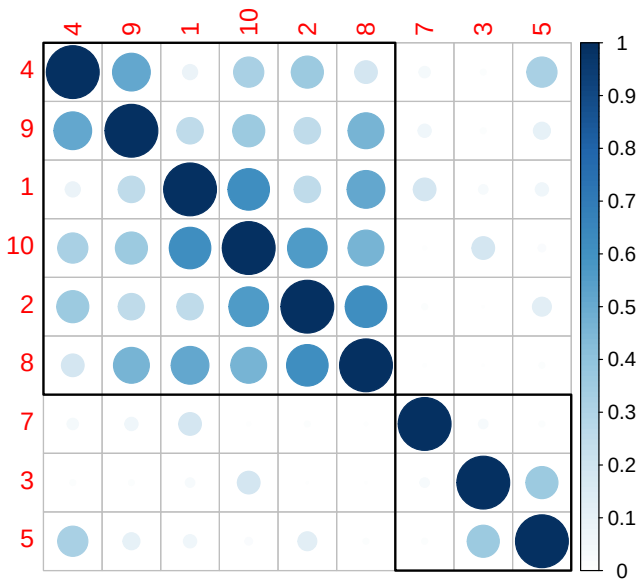

Figure 12: (Squared) correlations between syllable-level drummed annotations across drummers (indicated as numbers). The squares indicate the two main clusters of drummers.

2, 4, 8, 9, 10), and one cluster with 3 annotators (annotators 3, 5, 7). For data from each cluster, Random Forests were subsequently trained, using the same procedure and set of predictor variables as in chapter 4.1, but excluding the predictor variable ACCENT. We used the z-score normalized individual drumming performances of the various annotators present in the respective cluster as the dependent variable, i.e. no averaging was performed across the annotators present in each cluster.

The output was analyzed by ranking the importance of the predictors and further evaluated by checking the predictive accuracy of the resulting models.

5.2. Results

The Random Forest Model for Cluster 1 revealed an importance ranking dominated by POS, followed by acoustic prominence and syllable duration, again followed by the syllable's stressability and lastly the acoustic prominence of the previous syllable (cf. Figure 13). This ranking closely mirrors the results achieved by the expert annotators reported in the previous section. The Random Forest for the second cluster reveals a strong predominance of

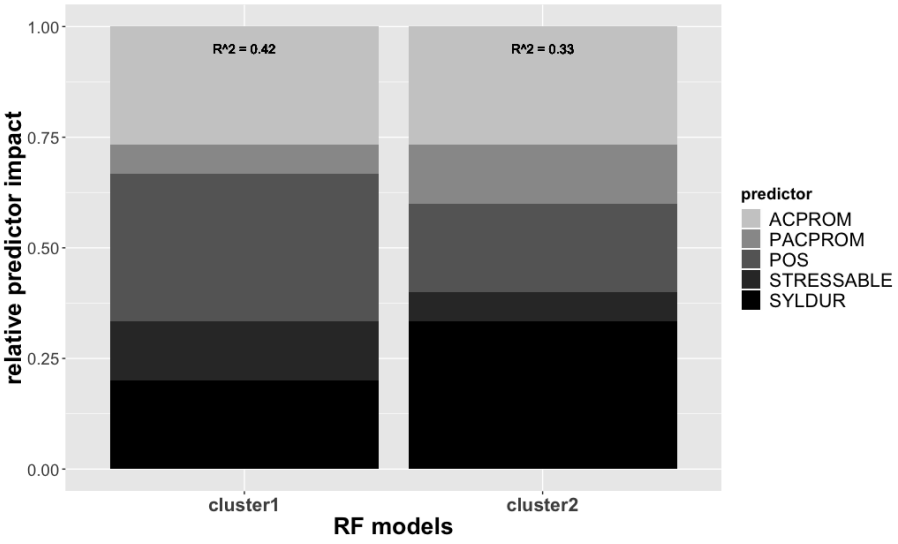

Figure 13: Importance ranking of predictor variables in Random Forest Models predicting syllable-level prominence annotation for clusters of drummers. For each model, predictive accuracies are given as RMSE (R^2)

signal variables, all of which outrank the structural variables (cf. Figure 13). However, in comparison to the Random Forest trained on the median performance across drummers, the predictive accuracy of the models is considerably lower, with $R^2 = 0.42$ for cluster 1, and only $R^2 = 0.33$ for cluster 2.

5.3. Discussion

While the mean drumming profiles revealed a predominance of signal-related cues in prominence drumming, a closer look at individual annotation patterns showed that these may indeed be underlying different strategies. Among our annotators, the larger group integrated structural and signal cues to prominence based on a strategy reflecting that of expert annotators — albeit less systematically, as the noise in the model reveals. The second group showed a behavior mostly mirroring the signal cues to prominence, but paid little attention to linguistic cues. It is now unclear, whether this attention to signal cues results from a misunderstanding of the task, and does not reflect these listeners' processing of prosody in realistic communicative situations, or whether they indeed perceive and process prosody in a different, more signal-oriented way. However, given the low predictive accuracy of the

Random Forest Model, we refrain from further speculation at this point. Obviously, more data is needed to conclusively answer this question.

6. General Discussion

Our analysis tried to shed light on several issues: First, we aimed at finding out whether a gesture-based account provides a viable alternative to the existing, established methods of prominence annotation. More specifically, we were looking for an annotation method that is accessible to naïve annotators, reveals fine-grained impressions of prosodic prominence for individual listeners and works without the simultaneous presentation of orthographic material. Both our feasibility study and our subsequent comparison of the gesture-based annotations with more conventional methods revealed that the gesture-based approach is generally suitable for gathering meaningful and fine-grained prominence impressions from naïve listeners close to real-time. Also, we could confirm that — while none of the different annotation methods delivers identical results — they all provide us with highly corresponding observations. Thus, we can basically continue comparing insights across the various methodological approaches. We have furthermore provided evidence supporting the hypothesis that perceptual impressions can indeed be transformed into motor movements encoding rich prosodic structures related to both structural and signal related cues (cf. section 1.5).

Secondly, we wanted to find out whether our gesture-based annotations reveal potential differences in integrating signal or structural cues to prosodic prominence. This general interest was extended to an analysis of whether the different annotation protocols, or the different domains of prominence annotation in themselves, trigger different cue integrations. When annotating word-level prominences, all annotation procedures appear to be comparatively stronger affected by structural linguistic considerations, probably cued by a shift in attention.

It was striking that even without the presence of orthographic input during the annotation phase, drummed word prominence annotations triggered high-level concepts to a similar degree as the other approaches. This can be interpreted in the sense that any attention to a higher order linguistic unit such as the word causes a conscious metalinguistic analysis that might strengthen the impact of structural cues in prominence processing.

For syllable-level annotations, the different annotation protocols reveal significant differences: Here, expert judgments appear to be affected stronger

by linguistic considerations than naïve listeners' judgments. This finding further strengthens our general scepticism against a reliance on expert annotations only, as these are apparently driven by theoretical preconceptions that may not perfectly reflect what non-linguists really do, or what anybody does in real life and real-time interactions. We also found that when comparing all annotation protocols, the gesture-based approach was the most signal-driven. We therefore definitely can say that the annotation method itself appears to influence the way that prominence cues are integrated. This should be considered in further investigations. Lastly, we were interested in whether there are unique strategies of prominence processing. Here, our analysis replicates findings by Baumann and Winter (2018), namely that there may be listeners paying more attention to an integration of structural and signal cues, while others may rely more strongly on signal-level cues. However, this insight should be treated as a hypothesis that needs further empirical investigation. Naturally, our analyses are heavily constrained by the linguistic and signal predictors taken into account. Structural cues of potentially high relevance such as word frequency or information structure were not taken into account. This was done for two reasons: As our data set was limited, we did not expect that frequency effects may become highly effective, and as our data set contained isolated sentences rather than contextually embedded utterances, information structural expectations could not be controlled. After now having shown the general feasibility of our gesture-based annotation, we are positive to conduct further investigations that consider larger, and — from a linguistic point of view — more interesting data. Here, our expectation would be that listener strategies of prominence perception may be dynamically adjustable to pragmatic needs (Bishop, 2016; Turnbull et al., 2017; Watson, 2010). We therefore hope to extend our annotation protocol to conversational data, where the pragmatic skills of the annotators may be triggered more strongly — and which are more in line with the demands of assessing prominence in daily communication. Along these lines, it would be very interesting to develop our method from a "reproduction task" into a real-time shadowing of the motor movements in listening or even conversational settings, perhaps by assessing subtle manual movements using motion capture or acceleration sensors. Such approaches would likely be able to shed light on the potential link between entrainment processes and speech perception.

References

Al Moubayed, S., Beskow, J., Granström, B., 2009. Auditory visual prominence. Journal on Multimodal User Interfaces 3 (4), 299–309.

Andreeva, B., Barry, W., Wolska, M., 2012. Language differences in the perceptual weight of prominence-lending properties. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association. Portland, OR, USA, pp. 2426–2429.

Arnold, D., Möbius, B., Wagner, P., 2011a. Comparing word and syllable prominence rated by naive listeners. In: Cosi, P., Mori, R. D., Fabbrizio, G. D., Pieraccini, R. (Eds.), INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association. Florence, Italy, pp. 1877–1880.

Arnold, D., Wagner, P., Baayen, H., 2013. Using Generalized Additive Models and Random Forests to Model Prosodic Prominence in German. In: Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., Perrier, P. (Eds.), INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association. Lyon, France, pp. 272–276.

Arnold, D., Wagner, P., Möbius, B., 2011b. Evaluating different rating scales for obtaining judgments of syllable prominence from naïve listeners. In: Proceedings of the 17th International Congress of Phonetic Sciences. Hong Kong, pp. 252–255.

Arnold, D., Wagner, P., Möbius, B., 2012. Obtaining prominence judgments from naïve listeners — Influence of rating scales, linguistic levels and normalisation. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association. pp. 2394–2397.

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.

Barry, W., Andreeva, B., Steiner, I., 2007. The Phonetic Exponency in Phrasal Accentuation in French and German. In: INTERSPEECH 2007,

8th Annual Conference of the International Speech Communication Association. Antwerp, Belgium, pp. 1010–1013.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67 (1), 1–48.

Baumann, S., Grice, M., Steindamm, S., 2006. Prosodic marking of focus domains-categorical or gradient? In: Proceedings of Speech Prosody 2006. Dresden, Germany, pp. 301–304.

Baumann, S., Riester, A., 2013. Coreference, Lexical Givenness and Prosody in German. Lingua 136, 16–37.

Baumann, S., Röhr, C. T., 2015. The perceptual prominence of pitch accent types in German. In: Proceedings of the 18th International Congress of the Phonetic Sciences. Glasgow, Scotland.

Baumann, S., Winter, B., 2018. What makes a word prominent? Predicting untrained German listeners' perceptual judgments. Journal of Phonetics 70, 20–38.

Beier, E. J., Ferreira, F., 2018. The Temporal Prediction of Stress in Speech and Its Relation to Musical Beat Perception. Frontiers in Psychology 9 (431).

Bell, A., Brenier, J. M., Gregory, M., Girand, C., Jurafsky, D., 2009. Predictability effects on durations of content and function words in conversational english. Journal of Memory and Language 60 (1), 92–111.

Bishop, J., 2016. Individual differences in top-down and bottom-up prominence perception. In: Barnes, J., Brugos, A., Shattuck-Hufnagel, S., Veilleux, N. (Eds.), Proceedings of Speech Prosody 2016. Boston, MA, USA, pp. 668–672.

Bishop, J., Kuo, G., Kim, B., in press. Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from rapid prosody transcription, *Journal of Phonetics*.

Boersma, P., Weenink, D., 2019. Praat: doing phonetics by computer [Computer program]. Version 6.0.56, retrieved 20 June 2019.
URL `http://www.praat.org/`

Bolton, T., 1894. Rhythm. American Journal of Psychology 6, 154–238.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Calhoun, S., 2010. How does Informativeness affect Prosodic Prominence. Language and Cognitive Processes 25 (7–9), 1099–1140.

Campbell, N., Beckman, M., 1997. Stress, Prominence and Spectral Tilt. In: Botinis, A., Kouroupetroglou, G., Carayiannis, G. (Eds.), Intonation: Theory, Models and Applications (Proceedings of an ESCA Workshop). ESCA and University of Athens Department of Informatics, Athens, Greece, pp. 67–70.

Cangemi, F., Grice, M., 2016. The importance of a distributional approach to categoriality in autosegmental-metrical accounts of intonation. Laboratory Phonology 7, 9.

Cole, J., Hualde, J. I., Smith, C. L., Eager, C., Mahrt, T., de Souza, R. N., 2019. Sound, structure and meaning: The bases of prominence ratings in english, french and spanish. Journal of Phonetics 75, 113,147.

Cole, J., Mahrt, T., Hualde, J. I., 2014. Listening for sound, listening for meaning: Task effects on prosodic transcription. In: Campbell, N., Gibbon, D., Hirst, D. (Eds.), Proceedings of Speech Prosody 2014. Dublin, Ireland, pp. 859–863.

Cole, J., Mahrt, T., Roy, J., 2017. Crowd-sourcing prosodic annotation. Computer Speech & Language 45, 300–325.

Cole, J., Mo, Y., Hasegawa-Johnson, M., 2010. Signal-based and expectation based factors in the perception of prosodic prominence. Laboratory Phonology 1 (2), 425–452.

Cole, J., Shattuck-Hufnagel, S., 2016. New Methods for Prosodic Transcription: Capturing Variability as a Source of Information. Laboratory Phonology 7, 8.

Condon, W., 1974. Neonate movement is synchronized with adult speech: Interactional participation and language acquisition. Science 183, 99–101.

Cummins, F., 2011. Periodic and aperiodic synchronization in skilled action. Frontiers in Human Neurosciences 5 (170).

Ćwiek, A., Wagner, P., 2018. The Acoustic Realization of Prosodic Prominence in Polish: Word-level Stress and Phrase-level Accent. In: Klessa, K., Bachan, J., Wagner, A., Karpiński, M., Śledziński, D. (Eds.), Proceedings of Speech Prosody 2018. Poznań, Poland, pp. 922–926.

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.

Dilley, L. C., McAuley, J. D., 2008. Distal prosodic context affects word segmentation and lexical processing. Journal of Memory and Language 59 (3), 294–311.

Erickson, F., Shultz, J., 1982. The counselor as gatekeeper: Social interaction in interviews. Academic Press, New York.

Eriksson, A., Thunberg, G. C., Traunmüller, H., 2001. Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. In: EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event. Aalborg, Denmark, pp. 399–402.

Escudero, P., Hayes-Harb, R., Mitterer, H., 2008. Novel second-language words and asymmetric lexical access. Journal of Phonetics 36 (2), 345–360.

Esteve-Gibert, N., Prieto, P., 2014. Infants temporally coordinate gesture-speech combinations before they produce their first words. Speech Communication 57, 301–316.

Fant, G., Kruckenberg, A., 1989. Preliminaries to the study of Swedish prose reading and reading style. STL-QPSR 30 (2), 1–80.

Fant, G., Kruckenberg, A., Liljencrants, J., 2000. Acoustic-phonetic Analysis of Prominence in Swedish. In: Botinis, A. (Ed.), Intonation. Kluwer Academic Publisher, pp. 55—86.

Féry, C., Kügler, F., 2008. Pitch accent scaling on given, new and focused constituents in German. Journal of Phonetics 36, 680–703.

Fougeron, C., Keating, P., 1997. Articulatory strengthening at edges of prosodic domains. Journal of the Acoustical Society of America 101 (6), 3728–3740.

Fry, D. B., 1958. Experiments in the perception of stress. Language and speech 1 (2), 126–152.

Garner, M., Lemon, J., Fellows, I., Sing, P., 2012. irr: Various coefficients of interrater reliability and agreement. Harvard University, Cambridge, USA, r package version 0.8.41.
URL `http://CRAN.R-project.org/package=stargazer`

Gussenhoven, C., Rietveld, A., 1988. Fundamental frequency declination in Dutch: testing three hypotheses. Journal of Phonetics 16, 355–369.

Hallgren, K. A., 2012. Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods of Psychology 8 (1), 23–34.

Hanssen, J., Peters, J., Gussenhoven, C., 2008. Prosodic effects of focus in Dutch declaratives. In: Barbosa, P. A., Madureira, S., Reis, C. (Eds.), Proceedings of Speech Prosody 2008. Campinas, Brazil, pp. 609–612.

Heckmann, M., 2014. Prosodic, Spectral and Visual Features for the Discrimination of Prominent and Non-prominent Words. In: Jahrestagung für Akustik. Oldenburg, Germany, pp. 59–63.

Heuft, B., 1999. Eine prominenzbasierte Methode zur Prosodieanalyse und -synthese. Peter Lang, Frankfurt a. M., also doctoral dissertation, Universität Bonn, Germany.

House, D., Beskow, J., Granström, B., 2001. Timing and Interaction of Visual Cues for Prominence in Audiovisual Speech Perception. In: Dalsgaard, P., Lindberg, B., Benner, H., hua Tan, Z. (Eds.), EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event. Aalborg, Denmark, pp. 387–390.

Jensen, C., Tøndering, J., 2005. Choosing a scale for measuring perceived prominence. In: Interspeech'2005 – Eurospeech, 9th European Conference on Speech Communication and Technology. Lisbon, Portugal, pp. 2385–2388.

Kakouros, S., Räsänen, O., 2016. Perception of Sentence Stress in Speech Correlates With the Temporal Unpredictability of Prosodic Features. Cognitive Science 40, 1739–1774.

Kalinli, O., Narayanan, S., 2009. Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information. IEEE Transactions on Audio, Speech and Language Processing 17 (5), 1009–1024.

Kochanski, G., Grabe, E., Coleman, J., 2005. Loudness predicts prominence: fundamental frequency lends little. Journal of the Acoustical Society of America 118, 1038–1054.

Krivokapic, J., Tiede, M., Tyrone, M., 2015. A kinematic analysis of prosodic structure in speech and manual gestures. In: Proceedings of the 18th International Congress of the Phonetic Sciences. Glasgow, Scotland, pp. 425–452.

Krivokapić, J., Tiede, M. K., Tyrone, M. E., 2017. A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. Laboratory Phonology 8 (1), 3.

Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., Jannedy, S., Michalsky, J., Niebuhr, O., Peters, J., Ritter, S., Röhr, C. T., Schweitzer, A., Schweitzer, K., Wagner, P., 2015. DIMA – Annotation Guidelines for German Intonation. In: Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, Scotland, p. 317.

Kunter, G., Plag, I., 2007. What is compound stress? In: Trouvain, J., Barry, W. J. (Eds.), Proceedings of the 16th International Congress of the Phonetic Sciences. Saarbrücken, Germany, pp. 1005–1008.

Lacheret, A., Simon, A. C., Goldman, J., Avanzi, M., 2013. Prominence Perception and Accent Detection in French: From Phonetic Processing to Grammatical Analysis. Language Sciences 39, 95–106.

Leonard, T., Cummins, F., 2010. The temporal relation between beat gestures and speech. Language and Cognitive Processes 26, 1295–1309.

Liberman, M., Prince, A., 1977. On stress and linguistic rhythm. Linguistic Inquiry 8 (2), 249–336.

Llanes-Coromina, J., Prieto, P., Rohrer, P. L., 2018a. Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task. In: Klessa, K., Bachan, J., Wagner, A., Karpiński, M., Śledziński, D. (Eds.), Proceedings of Speech Prosody 2018. Poznań, Poland, pp. 498–502.

Llanes-Coromina, J., Vilà-Giménez, I., Kushch, O., Borràs-Comes, J., Prieto, P., 2018b. Beat gestures help preschoolers recall and comprehend discourse information. Journal of Experimental Child Psychology 172, 168–188.

Loehr, D., 2004. Gesture and Intonation. Ph.D. thesis, Georgetown University, Washington, DC.

Loehr, D., 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. Laboratory Phonology 3 (1), 71–89.

Luchkina, T., Cole, J., 2014. Structural and Prosodic Correlates of Prominence in Free Word Order Language Discourse. In: Campbell, N., Gibbon, D., Hirst, D. (Eds.), Proceedings of Speech Prosody 2014. Dublin, Ireland, pp. 1119–1123.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2015. cluster: Cluster Analysis Basics and Extensions. R package version 2.0.1 — For new features, see the 'Changelog' file (in the package source).

Mahrt, T., Cole, J., Fleck, M., 2012. F0 and the Perception of Prominence. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association. Portland, OR, USA, pp. 2422–2425.

Malisz, Z., Berthelsen, H., Beskow, J., Gustafson, J., 2017. Controlling Prominence Realisation in Parametric DNN-Based Speech Synthesis. In: INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden, pp. 1079–1083.

Malisz, Z., Ćwiek, A., Wagner, P., 2015. The perception of prominence by Polish native speakers: a crowdsourcing study. In: Poznań Linguistics Meeting 2015.

Malisz, Z., Wagner, P., 2012. Acoustic-phonetic realisation of Polish syllable prominence: a corpus study. Vol. 14/15 of Special Issue of Speech and Language Technology. Polskie Towarzystwo Fonetyczne, Poznań, pp. 105–114.

Mendoza-Denton, N., Jannedy, S., 2011. Semiotic Layering Through Gesture and Intonation: A Case Study of Complementary and Supplementary Multimodality in Political Speech. Journal of English Linguistics 39 (3), 265–299.

Mooshammer, C., 2010. Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal effort in German. Journal of the Acoustical Society of America 127 (2), 1047–1058.

Ní Casaide, A., Yanushevskaya, I., Kane, J., Gobl, C., 2013. The Voice Prominence Hypothesis: the Interplay of F0 and Voice Source Features in Accentuation. In: Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., Perrier, P. (Eds.), INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association. Lyon, France, pp. 3527–3531.

Niebuhr, O., 2008. Interpretation of pitch patterns and its effects on accentual prominence in German. In: Proceedings of 3rd Conference on Tone and Intonation in Europe (TIE3). Lisbon, Portugal.

Niebuhr, O., 2009. F0-based rhythm effects on the perception of local syllable prominence. Phonetica 66 (1-2), 95–112.

Oh, M., Byrd, D., in press. Syllable-internal corrective focus in korean, *Journal of Phonetics*.

Parrell, B., Goldstein, L., Lee, S., Byrd, D., 2014. Spatiotemporal coupling between speech and manual motor actions. Journal of Phonetics 42, 1–11.

Phillips-Silver, J., Trainor, L. J., 2005. Feeling the beat: movement influences infant rhythm perception. Science 308 (5727), 1430.

Phillips-Silver, J., Trainor, L. J., 2007. Hearing what the body feels: Auditory encoding of rhythmic movement. Cognition 105 (3), 533–546.

Pitrelli, J. F., Beckman, M. E., Hirschberg, J., 1994. Evaluation of a prosodic transcription labeling reliability in the ToBI framework. In: Third International Conference on Spoken Language Processing (ICSLP 94). Yokohama, Japan, pp. 123–126.

Portele, T., Heuft, B., Widera, C., Wagner, P., Wolters, M., 2000. Perceptual Prominence. In: Sendlmeier, W. (Ed.), Speech and Signals. Aspects of Speech Synthesis and Automatic Speech Recognition. Hektor, Frankfurt a. M., pp. 97–115, festschrift for Wolfgang Hess on the occasion of his 60th birthday.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org/`

Repp, B. H., 2005. Sensorimotor synchronization: A review of the tapping literature. Psychonomic Bulletin & Review 12 (6), 969–992.

Richardson, C., Dale, R., Shockley, K., 2007. Synchrony and swing in conversation: Coordination, temporal dynamics, and communication. Oxford University Press, Oxford, pp. 75–94.

Roettger, T., Gordon, M., 2017. Methodological issues in the study of word stress correlates. Linguistics Vanguard 3 (1).

Rosenberg, A., Cooper, E., Levitan, R., Hirschberg, J., 2012. Cross-Language Prominence Detection. In: Ma, Q., Ding, H., Hirst, D. (Eds.), Proceedings of Speech Prosody 2012. Shanghai, China, pp. 278–281.

Samlowski, B., 2016. The syllable as a processing unit in speech production: evidence from frequency effects on coarticulation. Ph.D. thesis, Bielefeld University, Bielefeld, Germany.

Smith, R., Rathcke, T., in press. Dialectal phonology constrains the phonetics of prominence, *Journal of Phonetics*.

Sotschek, J., 1984. Sätze für Sprachgütemessungen und ihre phonologische Anpassung an die Deutsche Sprache. In: Tagungsband DAGA: Fortschritte der Akustik. Darmstadt, Germany, pp. 873–876.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics 8 (25).

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 (4), 323—-348.

Swerts, M., Krahmer, E., 2008. Facial expression and prosodic prominence: Effects of modality and facial area. Journal of Phonetics 36, 219—238.

Szalontai, Á., Wagner, P., Mády, K., Windmann, A., 2016. Teasing apart lexical and sentence accent in Hungarian and German. In: Draxler, C., Kleber, F. (Eds.), Proceedings of Phonetik und Phonologie im deutschsprachigen Raum (P und P) 12. München, Germany, pp. 215–218.

Tamburini, F., 2006. Reliable Prominence Identification in English Spontaneous Speech. In: Proceedings of Speech Prosody 2006. Dresden, Germany.

Tamburini, F., Wagner, P., 2007. On Automatic Prominence Detection for German. In: INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association. Antwerp, Belgium, pp. 1809–1812.

Tanenhaus, M. K., Flanigan, H. P., Seidenberg, M. S., 1980. Orthographic and phonological activation in auditory and visual word recognition. Memory & Cognition 8 (6), 513–520.

Taylor, P. A., 2000. Analysis and Synthesis of Intonation using the TILT Model. Journal of the Acoustical Society of America 107 (3), 1697—1714.

Terken, J., 1991. Fundamental frequency and perceived prominence of accented syllables. Journal of the Acoustical Society of America 89 (4), 1768–1776.

Tkachman, O., Hall, K., Fuhrman, R., Aonuki, Y., in press. Visible amplitude: Towards a quantifying prominence in sign language, *Journal of Phonetics*.

Turco, G., Dimroth, C., Braun, B., 2013. Intonational Means to Mark Verum Focus in German and French. Language and Speech 4 (56), 461–491.

Turk, A. E., Sawusch, J. R., 1996. The processing of duration and intensity cues to prominence. Journal of the Acoustical Society of America 99, 3782–3790.

Turnbull, R., Royer, A. J., Ito, K., Speer, S. R., 2017. Prominence perception is dependent on phonology, semantics, and awareness of discourse. Language, Cognition and Neuroscience 32 (8), 1017–1033.

Vainio, M., Järvikivi, J., 2006. Tonal features, intensity, and word order in the perception of prominence. Journal of Phonetics 34 (3), 319–342.

Vogel, R., van de Vijver, R., Kotz, S., Kutscher, A., Wagner, P., 2015. Function words in rhythmic optimisation. Vol. 286 of Trends in Linguistics. Mouton de Gruyter, Berlin, Ch. 10, pp. 255–276.

Wagner, P., 2005. Great Expectations – Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates. In: Interspeech'2005 – Eurospeech, 9th European Conference on Speech Communication and Technology. Lisbon, Portugal, pp. 2381–2384.

Wagner, P., Bryhadyr, N., 2017. Mutual Visibility and Information Structure Enhance Synchrony between Speech and Co-Speech Movements. Journal of Multimodal Communication Studies 4 (1-2), 69–74.

Wagner, P., Malisz, Z., Inden, B., Wachsmuth, I., 2013. Interaction Phonology – A temporal co-ordination component enabling representational alignment within a model of communication. John Benjamins, Amsterdam, pp. 109–132.

Wagner, P., Malisz, Z., Kopp, S., 2014. Speech and Gesture in Interaction: An Overview. Speech Communication 57, 209–232.

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., Ní Chasaide, A., Niebuhr, O., Rousier-Vercruyssen, L., Simon, A. C., Simko, J., Tesser, F., Vainio, M., 2015a. Different Part of the Same Elephant: A Roadmap to Disentangle and Connect Different Perspectives on Prosodic Prominence. In: Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, Scotland.

Wagner, P., Trouvain, J., Zimmerer, F., 2015b. In defense of stylistic diversity in speech research. Journal of Phonetics 48, 1–12.

Walker, J., 2008. MIDICSV: Convert MIDI files to and from CSV. URL `http://www.fourmilab.ch/webtools/midicsv/`

Wang, D., Narayanan, S., 2007. An Acoustic Measure for Word Prominence in Spontaneous Speech. IEEE Transactions on Audio, Speech and Language Processing 15 (2), 690–701.

Watson, D. G., 2010. Chapter 4 - The Many Roads to Prominence: Understanding Emphasis in Conversation. In: Ross, B. H. (Ed.), The Psychology of Learning and Motivation. Vol. 52 of Psychology of Learning and Motivation. Academic Press, pp. 163–183.

Watson, D. G., Arnold, J. E., Tanenhaus, M. K., 2008. Tic tac TOE: Effects of predictability and importance on acoustic prominence in language production. Cognition 106, 1548–1557.

Wei, T., 2013. corrplot: Visualization of a correlation matrix. R package version 0.73.
URL `http://CRAN.R-project.org/package=corrplot`

Widera, C., Portele, T., Wolters, M., 1997. Prediction of word prominence. In: EUROSPEECH '97 – 5th European Conference on Speech Communication and Technology. Vol. 2. Rhodes, Greece, pp. 999–1002.

Wightman, C. W., 1993. Perception of multiple levels of prominence in spontaneous speech. The Journal of the Acoustical Society of America 94 (3), 1881–1881.

Windmann, A., Wagner, P., Tamburini, F., Arnold, D., Oertel, C., 2010. Automatic Prominence Annotation of a German Speech Synthesis Corpus: Towards Prominence-Based Prosody Generation for Unit Selection Synthesis. In: Sagisaka, Y., Tokuda, K. (Eds.), Proceedings of the 7th ISCA Tutorial and Research Workshop on Speech Synthesis. pp. 377–382.

Xu, Y., 1999. Effects of tone and focus on the formation and alignment of f0 contours. Journal of Phonetics 27 (2), 55–105.

Yun, K., Watanabe, K., Shimojo, S., 2012. Interpersonal body and neural synchronization as a marker of implicit social interaction. Scientific Reports 2, 959.

Appendix A. Sentence material used in the annotation

1. Frische Gardinen hängen am Fenster.
2. Der junge Zugbegleiter pfeift zur Abfahrt.
3. Rückt die Stühle an den Tisch!
4. Es ist acht Uhr morgens.
5. “Zug endet hier” verkündet die Ansage.
6. Die Bremsen quietschen grässlich.
7. Die Nacht haben Meiers gut geschlafen.
8. Die Fahrt war ja mächtig kurz.
9. Wir hören den plätschernden Bach.
10. Mutter konnte länger schlafen.
11. Es gehört zu einer Feldscheune.
12. Riecht ihr nicht die frische Luft?
13. Ob ich Süßigkeiten kaufen darf?
14. Über die Felder weht ein Wind.
15. Da möchte ich gerne mit.
16. Zieht vielleicht die festen Schuhe an!
17. Jetzt suche ich das Weißbrot

18. Da läuft der Zug ein.

19. Die Ärzte sind damit gar nicht einverstanden.

20. Aus dem Radio klingt Musik.

Appendix B. LMEMs testing impact of individual items on drumming impact forces

	Estimate	Std. Error	df	t value
(Intercept)	71.47	1.57	25.34	45.54
6s	2.51	1.32	305.61	1.91
aI	2.81	1.35	258.39	2.08
aIn	2.26	1.08	137.01	2.10
am	-0.50	1.19	843.28	-0.42
an	2.03	1.10	152.73	1.85
ap	4.15	1.40	127.33	2.96
aUs	2.40	1.66	63.04	1.45
axt	7.37	1.39	138.00	5.31
b@	2.79	1.28	284.35	2.18
b@n	-1.30	1.26	565.05	-1.04
b6	1.37	1.33	234.86	1.03
bax	6.02	1.20	179.78	5.00
brEm	3.00	1.29	471.82	2.33
bro:t	1.74	1.33	184.41	1.31
d@n	1.80	1.20	80.06	1.51
d@t	0.91	1.17	272.84	0.78
d6	5.56	1.39	126.64	4.01
da:	3.71	1.04	444.65	3.58
darf	4.37	1.32	89.62	3.30
das	3.01	1.38	206.02	2.19
de:6	0.66	1.08	493.48	0.61
de:m	0.31	1.68	64.61	0.19
de:n	0.91	1.13	126.86	0.80
di:	1.13	0.90	944.71	1.25
djo	1.94	1.68	64.61	1.15

E:6ts	3.88	1.22	662.93	3.17
en	2.10	1.24	868.60	1.70
Es	2.53	1.17	205.54	2.16
f@n	2.50	1.06	321.86	2.37
fa:rt	2.03	1.13	533.39	1.79
fE6	3.47	1.14	251.54	3.04
fEl	7.05	1.29	366.36	5.45
fElt	7.33	1.27	231.78	5.79
fEns	3.73	1.32	113.90	2.84
fEs	7.15	1.16	308.11	6.19
fi:	-1.21	1.09	1661.35	-1.10
frI	2.81	1.08	351.08	2.61
g@	0.15	1.07	477.13	0.14
g@ns	1.80	1.40	92.75	1.28
ga:r	5.48	1.25	201.70	4.39
gar	-1.55	1.24	683.76	-1.25
gEr	4.64	1.29	296.51	3.60
glaI	0.80	1.30	237.28	0.61
grEs	6.52	1.32	309.23	4.93
gu:t	2.21	1.27	212.75	1.73
h2:	3.73	1.28	252.88	2.92
h2:6t	3.09	1.33	221.91	2.32
ha:	2.46	1.26	528.09	1.95
hE	1.36	1.37	145.22	0.99
hi:6	3.23	1.26	437.75	2.56
i:6	-0.82	1.26	532.57	-0.65
IC	-0.08	1.06	457.98	-0.08
Ist	0.20	1.35	182.22	0.15
ja:	0.82	1.31	274.20	0.62
jEtst	3.70	1.29	326.98	2.88
jU	1.42	1.26	536.62	1.13
kaI	1.85	1.31	252.12	1.42
kaU	4.68	1.31	233.24	3.57
klINt	4.97	1.71	65.91	2.90
kOn	0.94	1.32	251.16	0.72
kUrts	2.26	1.41	284.47	1.61
kvi:	3.84	1.29	444.72	2.98
kYn	2.01	1.29	281.69	1.56

l@	1.32	1.28	398.90	1.03
laICt	-0.30	1.15	425.46	-0.26
lE	7.78	1.34	212.70	5.81
lIC	3.23	1.32	390.32	2.45
lOYft	-1.38	1.24	642.20	-1.11
lUft	3.95	1.42	223.57	2.78
m9C	0.15	1.24	651.36	0.12
maI	5.63	1.28	598.98	4.40
mEC	5.69	1.24	342.41	4.59
mIt	3.11	1.10	234.62	2.83
mOr	4.78	1.34	144.74	3.55
mu	2.79	1.72	55.83	1.63
mU	6.24	1.21	456.29	5.14
n@	0.88	1.15	195.70	0.76
N@	0.25	1.30	296.94	0.20
n@n	0.71	1.31	235.20	0.54
N@n	-1.65	1.32	304.27	-1.24
n6	3.68	1.39	186.75	2.65
N6	3.01	1.34	237.96	2.25
naxt	1.73	1.27	499.64	1.37
nICt	2.00	1.09	555.15	1.84
Op	2.04	1.22	394.19	1.67
pfaIft	5.89	1.21	616.10	4.88
plE	7.04	1.34	139.76	5.27
r@n	1.35	1.31	171.62	1.03
ra:	5.53	1.69	64.70	3.27
ri:Ct	4.14	1.23	673.74	3.37
rYkt	5.22	1.23	430.60	4.23
S@	0.12	1.12	182.33	0.11
sIC	2.47	1.27	342.35	1.94
Sla:	2.27	1.18	252.80	1.92
SOY	3.73	1.16	335.37	3.23
Stan	1.98	1.31	145.11	1.51
Sty:	6.08	1.28	323.16	4.75
Su:	3.35	1.10	236.08	3.05
t@	-0.21	1.11	289.42	-0.19
t@n	1.97	1.10	192.39	1.80
t6	1.84	1.04	330.73	1.78

tIC	-0.20	1.30	140.34	-0.15
tIS	4.53	1.41	182.95	3.22
tS@n	2.00	1.31	425.92	1.52
tS6n	3.38	1.40	76.80	2.41
tsi:t	3.39	1.08	1113.04	3.15
tsu:	1.18	1.30	332.08	0.91
tsu:6	0.16	1.35	92.17	0.12
tsu:k	4.19	1.06	300.15	3.94
u:6	3.16	1.41	127.84	2.24
va:r	2.12	1.28	473.11	1.66
vaIs	5.90	1.37	120.42	4.32
ve:t	4.59	1.33	209.38	3.45
vi:6	2.53	1.27	192.15	1.99
vInt	5.53	1.30	294.08	4.25
x@	-1.02	1.37	234.97	-0.75
y:	2.24	1.26	339.86	1.78
z@n	0.29	1.32	311.67	0.22
za:	1.41	1.39	132.25	1.02
zi:k	4.18	1.68	60.44	2.49
zInt	1.40	1.23	562.65	1.13
zu:	-0.48	1.35	240.15	-0.36
zy:	6.45	1.28	258.09	5.03

Table B.4: LMEM predicting drummed impact forces from individual syllable items.

	Estimate	Std. Error	df	t value
(Intercept)	68.55	1.99	13.48	34.52
aIn6	-0.58	1.67	31.30	-0.35
am	0.41	1.59	51.14	0.26
an	0.33	1.35	28.52	0.24
apfa:rt	3.86	1.63	21.71	2.37
aUs	-0.93	1.61	21.19	-0.58
axt	4.10	1.63	32.50	2.51
bax	4.44	1.47	35.56	3.01
brEmz@n	2.74	1.60	30.31	1.71
da:	1.70	1.22	49.46	1.40

da:mIt	-0.80	1.84	68.98	-0.43
darf	2.88	1.43	39.69	2.01
das	-1.69	1.64	35.53	-1.04
de:6	-2.09	1.19	49.28	-1.76
de:m	-1.18	1.69	19.06	-0.70
de:n	-1.40	1.20	72.80	-1.16
di:	-1.39	1.01	37.36	-1.38
E:6tst@	6.03	1.86	55.81	3.24
end@t	2.38	1.66	25.72	1.43
Es	-1.29	1.36	30.43	-0.95
f@ng@Sla:	3.55	1.64	25.87	2.16
fa:rt	3.55	1.62	31.79	2.18
fE6kYnd@t	1.25	1.69	22.17	0.74
fE6Stand@naIn	4.66	1.90	27.52	
fEld6	4.36	1.41	65.82	3.09
fEltSOYn@	5.59	1.48	29.40	3.77
fEst@n	1.98	1.66	32.32	1.20
fi:laICt	2.35	1.63	39.50	1.44
frIS@	1.95	1.17	63.23	1.67
g@h2:6t	2.65	1.67	30.43	1.58
ga:r	3.47	1.68	108.84	2.07
gardi:n@n	5.70	1.62	37.50	3.51
gErn@	1.70	1.61	48.43	1.05
grEslIC	3.52	1.59	31.72	2.22
gu:t	0.90	1.42	59.07	0.64
h2:r@n	4.74	1.60	53.97	2.96
ha:b@n	0.35	1.60	41.54	0.22
hEN@n	1.34	1.59	42.74	0.84
hi:6	3.40	1.62	34.41	2.10
i:6	-0.95	1.57	37.28	-0.60
IC	-0.63	1.27	31.43	-0.50
Ist	-1.21	1.66	29.60	-0.73
ja:	-0.36	1.61	33.31	-0.22
jEtst	2.77	1.55	53.37	1.78
jUN@	1.46	1.60	32.08	0.92
kaUf@n	2.00	1.62	36.33	1.24
klINt	-0.48	1.69	19.69	-0.29
kOnt@	-0.17	1.69	18.69	-0.10

kUrts	1.94	1.63	25.12	1.19
kvi:tS@n	3.46	1.56	38.83	2.22
lEN6	3.81	1.66	19.20	2.30
lOYft	-2.05	1.40	36.86	-1.46
lUft	6.15	1.60	22.27	3.84
m9Ct@	-0.05	1.64	43.67	-0.03
maI6s	2.66	1.62	34.21	1.64
mECtIC	4.00	1.40	58.40	2.85
mIt	5.08	1.42	116.46	3.59
mOrg@ns	4.40	1.47	28.39	3.00
mUt6	4.43	1.60	20.74	2.77
muzi:k	2.63	1.50	14.92	1.75
naxt	1.67	1.64	32.63	1.02
nICt	-0.58	1.38	36.91	-0.42
Op	-0.04	1.54	50.81	-0.02
pfaIft	2.08	1.58	35.34	1.32
plEtS6nd@n	1.00	1.61	48.92	0.62
ra:djo	5.69	1.67	19.59	3.42
ri:Ct	3.13	1.50	71.86	2.09
rYkt	3.31	1.51	80.02	2.20
Sla:f@n	2.36	1.68	19.31	1.40
Sty:l@	5.78	1.56	50.10	3.69
Su:@	4.75	1.42	95.78	3.36
t6fEns	2.47	1.47	30.54	1.67
tIS	4.95	1.64	19.74	3.02
tsi:t	2.04	1.55	67.08	1.31
tsu:	0.41	1.65	33.43	0.25
tsu:6	-0.52	1.39	47.80	-0.37
tsu:k	1.30	1.24	26.61	1.05
tsu:kb@glaIt6	4.26	1.61	28.85	2.64
u:6	-1.88	1.66	30.44	-1.13
va:r	0.50	1.58	40.57	0.32
vaIsbro:t	3.61	1.44	43.86	2.50
ve:t	0.26	1.45	39.23	0.18
vi:6	-0.68	1.54	61.13	-0.44
vInt	4.63	1.48	20.92	3.14
y:b6	1.25	1.38	99.31	0.91
za:g@an	3.94	1.68	18.38	2.35

zInt	0.61	1.85	66.30	0.33
zu:x@	-1.25	1.66	31.50	-0.75
zy:sICkaIt@n	4.69	1.65	30.61	2.85

Table B.5: LMEM predicting drummed impact forces from individual word items.

Appendix C. LMEMs comparing word-level annotation protocols

	Estimate	Std. Error	t value
(Intercept)	20.47	1.33	15.35
p-scores	-0.05	0.45	-0.11
orth p-scores	-11.27	1.88	-5.98
drummed	0.05	0.12	0.44

Table C.6: LMEM predicting fine-grained expert annotations from cumulative naïve p-scores, orthographic "p-scores" and drummed impact forces.

	Estimate	Std. Error	t value
(Intercept)	0.04	0.02	1.74
expert	0.00	0.00	2.57
orth p-scores	0.32	0.04	7.71
drummed	0.06	0.01	5.40

Table C.7: LMEM predicting cumulative naïve p-scores from fine-grained expert annotations, orthographic "p-scores" and drummed impact forces.

	Estimate	Std. Error	t value
(Intercept)	0.47	0.05	9.87
expert	-0.01	0.00	-8.03
p-scores	0.03	0.01	2.90
drummed	-0.01	0.00	-2.25

Table C.8: LMEM predicting cumulative orthographic "p-scores" from fine-grained expert annotations, p-scores and drummed impact forces.

	Estimate	Std. Error	t value
(Intercept)	-0.94	0.13	-7.37
experts	0.04	0.01	4.66
orth pscores	0.52	0.24	2.18
p-scores	1.40	0.22	6.23

Table C.9: LMEM predicting drummed impact forces from fine-grained expert annotations, cumulative orthographic "p-scores" and cumulative naïve p-scores and drummed impact forces.

Appendix D. LMEMs comparing syllable-level annotation protocols

	Estimate	Std. Error	df	t value
(Intercept)	6.88	0.63	7.57	11.00
pscores	10.95	1.45	481.21	7.53
orth pscores	10.73	1.02	482.36	10.54
drummed	2.67	0.44	481.86	6.09

Table D.10: LMEM predicting fine-grained expert annotations from cumulative naïve p-scores, orthographic "p-scores" and drummed impact forces.

	Estimate	Std. Error	df	t value
(Intercept)	-0.06	0.02	79.60	-3.18
experts	0.02	0.00	479.32	10.66
p-scores	0.40	0.06	483.01	6.83
drummed	0.07	0.02	480.60	4.17

Table D.11: LMEM predicting cumulative orthographic "p-scores" from fine-grained expert annotations, p-scores and drummed impact forces.

	Estimate	Std. Error	df	t value
(Intercept)	0.07	0.02	11.49	3.90
experts	0.01	0.00	470.89	7.46
orth pscores	0.21	0.03	489.20	6.69
drummed	0.11	0.01	474.11	8.44

Table D.12: LMEM predicting cumulative orthographic "p-scores" from fine-grained expert annotations, p-scores and drummed impact forces.

	Estimate	Std. Error	df	t value
(Intercept)	-0.64	0.05	13.97	-14.01
percProm	0.03	0.00	484.50	5.92
orth_ratings	0.46	0.11	487.90	4.15
pscores	1.19	0.14	487.13	8.40

Table D.13: LMEM predicting drummed impact forces from fine-grained expert annotations, cumulative orthographic "p-scores" and cumulative naïve p-scores and drummed impact forces.

Appendix E. Analyses scripts

Appendix E.1. R-script for the analysis of global variance in drummed prominences with linear mixed models, to see whether there is a unique impact of linguistic structure (words, syllables) on drumming force

```
#read data file, containing drumming velocity, sentence
    information, word information, syllable information
    , speaker information, drummer information
# example header:
# drummer speaker word  wordnr  (syllable) syllnr
   sentence  velocity

data.syll=read.table("mysyll_drumming_results.txt",
    fill=TRUE, header=TRUE, sep="\t")
data.word=read.table("myword_drumming_results.txt",
    fill=TRUE, header=TRUE, sep="\t")

test_randomness_syll.lmer <- lmer(data=data, velocity ~
    syllable + (1|sentence) + (1|wordnr) + (1|syllnr) +
```

```
    (1|speaker) + (1|drummer))

test_randomness_word.lmer <- lmer(data=data, velocity ~
    word + (1|wordnr) + (1|sentence) + (1|speaker) +
   (1|drummer))
```

Appendix E.2. R-script for the analysis of global variance in drummed prominences with linear mixed models, to see whether there is a unique impact of linguistic structure (words, syllables) on drumming force

```
library(lme4)

#set path to working directory
setwd("/mypath/")

#read data file, containing drumming velocity, sentence
    information, word information, syllable information
   , speaker information, drummer information
# example header:
# drummer speaker word  wordnr  (syllable) syllnr
   sentence  velocity

data.syll=read.table("mysyll_drumming_results.txt",
   fill=TRUE, header=TRUE, sep="\t")
data.word=read.table("myword_drumming_results.txt",
   fill=TRUE, header=TRUE, sep="\t")

test_randomness_syll.lmer <- lmer(data=data, velocity ~
    syllable + (1|sentence) + (1|wordnr) + (1|syllnr) +
    (1|speaker) + (1|drummer))

test_randomness_word.lmer <- lmer(data=data, velocity ~
    word + (1|wordnr) + (1|sentence) + (1|speaker) +
   (1|drummer))

summary(test_randomness_syll.lmer)
```

```
summary(test_randomness_word.lmer)
```

Appendix E.3. R-script for comparing dependencies between various forms of prominence annotation using linear mixed models

```
library(lme4)

#set path to working directory

setwd("/mypath/")

#read data tables containing syllable or word
   annotations gathered with different annotation
   schemes
#Read data file
#explanation of relevant data table headers:
#satznummer: sentence id
#sprecher: speaker id
#wort:   transcribed word
#silbe: transcribed syllable
#mean_velocity: drumming force (z-scored by sentence),
   averaged across drummers
#velocity: individual drumming force (z-scored by
   sentence)
#orth: orthographic representation
#orthProm: pscores based on orthographic input
#percProm: expert prominence ratings (0-31 scale, means
   )
#pscores: p-scores based on cumulative prominence
   ratings

#expected file format:
#satznummer    sprecher   wort  mean_velocity   orth
      cumulProm   orthProm    percProm    pscores

#choose data frame for word or syllable-based
   annotations:
```

```
#comp.dat=read.table("comparison_syllannos_averages.txt
    ", fill=TRUE, header=TRUE, sep="\t")
comp.dat=read.table("comparison_wordannos_averages_korr
    .txt", fill=TRUE, header=TRUE, sep="\t")


#Build models with annotation method as fixed factor
    and words/sentences/speakers as random factors
#Determine if the various annotations can be predicted
    based on alternative annotation methods

percProm.lmer <- lmer(data=comp.dat,percProm ~ pscores
   + orthProm + mean_velocity + (1|sprecher) + (1|wort)
    + (1|satznummer))
orthProm.lmer <- lmer(data=comp.dat,orthProm ~ percProm
    + pscores + mean_velocity + (1|wort) + (1|
   satznummer) + (1|sprecher))
pscores.lmer <- lmer(data=comp.dat,pscores ~ percProm +
    orthProm + mean_velocity + (1|wort) + (1|satznummer
   ) + (1|sprecher))
velocities.lmer <- lmer(data=comp.dat,mean_velocity ~
   percProm + orthProm + pscores + (1|wort) + (1|
   satznummer) + (1|sprecher))


summary(percProm.lmer)
summary(orthProm.lmer)
summary(pscores.lmer)
summary(velocities.lmer)
```

Appendix E.4. R-Script for a assessing inter-annotator correspondence using global ICC and inter-annotator correlations between individual drummers

```
library(irr)
library(corrplot)
```

```
#Set path to your data files

setwd("/mypath/")

#Read data table
#expected data format is table with header containing
   the following columns
#satznummer: sentence number (unique sentence id)
#sprecher: speaker id
#mean velocity: average velocity across sentence
#velocity: mean velocity across drummers
#drummer1, drummer2, drummer3.... drummerN: individual
   drummer's drumming velocities (z-score normalized
   per sentence)
#orth_ratings: orthography-based p-scores
#ratings: p-scores
#percProm: expert prominences

#example:
#line   abbrev  satznummer      sprecher
   wortnummer      wort    mean_velocity   velocity
            orth    ratings orth_ratings    percProm
            drummer2        drummer4        drummer6
            drummer8        drummer10       drummer12
           drummer14       drummer16       drummer18


comp.dat=read.table("my_annotations.txt", fill=TRUE,
   header=TRUE, sep="\t")

#Calculate global constistency using IntraClass
   Correlations (consistency) across relevant drummers.
#For word-drummings columns 13-21 of the data table are
    used
#For syll drummings colums 24-32 of the data table are
   used
```

```
iccs = icc(comp.dat[,24:32], model="twoway", type="
    consistency", unit="average")


#check for different sentences and number of sentences
    and number of speakers

sentences <-  unique(comp.dat$satznummer)
nsentences <- length(unique(comp.dat$satznummer))
nspeakers <- length(unique(comp.dat$sprecher))

#prepare data frames for different speakers (f, l, m)

comp.dat_f <- comp.dat[comp.dat$sprecher == "f",]
comp.dat_l <- comp.dat[comp.dat$sprecher == "l",]
comp.dat_m <- comp.dat[comp.dat$sprecher == "m",]

#set matrix for correlation plot (for number of
    annotators, here: 9x9 for syllable annotations, and
    10x10 for word annotations)

#prev_matrix <- matrix(c(replicate(0,100)), 10, 10)
prev_matrix <- matrix(c(replicate(0,81)), 9, 9)

#Calculate correlations by-sentence/by-speaker per
    individual drummer
#For word-drummings columns 13-21 of the data table are
     used
#For syll drummings colums 24-32 of the data table are
    used

for(i in 1:nsentences){
  local.dat_f <- comp.dat[comp.dat$sprecher == "f" &
     comp.dat$satznummer == sentences[i],]
  local.dat_l <- comp.dat[comp.dat$sprecher == "l" &
     comp.dat$satznummer == sentences[i],]
  local.dat_m <- comp.dat[comp.dat$sprecher == "m" &
     comp.dat$satznummer == sentences[i],]
```

```
  local.correlations_f.cor = cor(local.dat_f[,24:32],
     method="spearman")

  local.correlations_l.cor = cor(local.dat_l[,24:32],
     method="spearman")

  local.correlations_m.cor = cor(local.dat_m[,24:32],
     method="spearman")

#add local correlations to list

  local_list <- list(local.correlations_f.cor,local.
     correlations_l.cor, local.correlations_m.cor)
  all_list <- list(prev_matrix, local_list)
  prev_matrix <- local_list
}

#Build an array of medians for plotting for 10x10 or 9
   x9 matrix

#arr <- array( unlist(all_list) , c(10,10,3) )
#different array for word annotations (9 annotators)
arr <- array(unlist(all_list) , c(9,9,3) )

mean_corrs <- apply(arr,1:2, median, na.rm=TRUE)

#make rownames and colnames for 9x9 or 10x10 matrix

rownames(mean_corrs) <- c("1", "2", "3", "4", "5", "7",
    "8", "9", "10")
colnames(mean_corrs) <- c("1", "2", "3", "4", "5", "7",
    "8", "9", "10")


#rownames(mean_corrs) <- c("1", "2", "3", "4", "5",
   "7", "8", "9")
#colnames(mean_corrs) <- c("1", "2", "3", "4", "5",
```

```
    "7", "8", "9")


#square mean correlation coefficients to avoid negative
    values (for cluster analysis);adapt to matrix size

mean_corrs <- (mean_corrs^2)[1:9,1:9]
mean_corrs.dist = dist(mean_corrs)

corrs_clust <- agnes(mean_corrs.dist)
pltree(corrs_clust)

#plot
corrplot(mean_corrs, order="hclust", addrect = 2, cl.
   lim=c(0,1))
```

Appendix E.5. Training Random Forests to determine influential factors on drumming velocities or other dependent variables capturing prosodic prominence

```
library(party)

#set working directory

setwd("/mypath/")

#select data frame containing drumming data; data
   should contain the dependent variable (e.g. drumming
    velocity, p-scores, expert prominences) and a set
   of numerical and categorical predictor variables (
   POS, accent, f0, duration)
#data example (header):
# participant sentence  speaker word  POS accent
   duration  intensity acoustic_prom f0  accentdist
   velocity  pscores mean_exp_proms  mean_velocity

data=read.table("myresults.txt", fill=TRUE, header=TRUE
   , sep="\t")
```

```
data <- na.omit(data)
data.new <- data

#Random Forest Building
#choose data set if selection is necessary (clusters or
    individual drummers), examples:

data_cluster <- data.new[ which(data.new$participant
   =="1" | data.new$participant=="2" | data.
   new$participant=="3"| data.new$participant=="4" |
   data.new$participant=="5"), ]

data_drummer7 <- data.new[ which(data.new$participant
   =="7"), ]


#Random Forests computed with party package

set.seed(1234)

#Splitting the data into training and test set

split.dat <- sample(2,nrow(data.new), replace=TRUE,
   prob=c(0.7,0.3))
traindata <- data.new[split.dat==1,]
testdata <- data.new[split.dat==2,]

#Set formula for regression (here: predict mean
   drumming velocities based on acoustic prominences,
   durations, distance to accent (in syllables), part
   of speech)

myFormula <- mean_velocity ~ ac_prom + duration +
   accent_dist + POS

#Train a Random Forest (adapt based on your needs and
   questions asked)
```

```
fitdrums <- cforest(myFormula, data = traindata,
    weights = NULL,  controls = cforest_unbiased(),
    xtrafo = ptrafo, ytrafo = ptrafo, scores = NULL)

#Weigh importance of predictor variables

varimp <- varimp(fitdrums)

#predict drums for RMSE calculation

predicted_drums <- predict(fitdrums, newdata = testdata
    , type="response",OOB = TRUE)

#calculate RMSE relative to y (here: mean drumming
    velocities)

testdata$predicted_velocity = predicted_drums
y <- testdata$mean_velocity
1-sum((y-predicted_velocity)^2)/sum((y-mean(y))^2)
```