Research Article

Nele Ots* and Piia Taremaa

A perceptual study of language chunking in Estonian

https://doi.org/10.1515/opli-2020-0182
received December 29, 2020; accepted November 14, 2021

Abstract: Two studies investigate the production and perception of speech chunks in Estonian. A corpus study examines to what degree the boundaries of syntactic constituents and frequent collocations influence the distribution of prosodic information in spontaneously spoken utterances. A perception experiment tests to what degree prosodic information, constituent structure, and collocation frequencies interact in the perception of speech chunks. Two groups of native Estonian speakers rated spontaneously spoken utterances for the presence of disjunctures, whilst listening to these utterances (N = 47) or reading them (N = 40). The results of the corpus study reveal a rather weak correspondence between the distribution of prosodic information and boundaries of the syntactic constituents and collocations. The results of the perception experiments demonstrate a strong influence of clause boundaries on the perception of prosodic discontinuities as prosodic breaks. Thus, the results indicate that there is no direct relationship between the semantico-syntactic characteristics of utterances and the distribution of prosodic information. The percept of a prosodic break relies on the rapid recognition of constituent structure, i.e. structural information.

Keywords: speech processing, sentence prosody, chunking, Estonian, rapid prosody transcription

1 Introduction

Utterances spoken in communicative contexts frequently diverge from the organization of written sentences or read-aloud speech (see e.g. Blaauw 1994, Schegloff 1996). Within their turns of speaking, speakers usually convey their ideas in spurts known as chunks of speech. There is evidence that speech chunks frequently correspond with syntactically defined and/or semantically related groupings of words (see e.g. Berkovits 1994, Fon et al. 2011, Schafer et al. 2000, 2005, Snedeker and Trueswell 2003). However, there is also evidence for opposite cases involving chunks containing multiple syntactically defined elements or elements that appear uncompleted in terms of syntax and semantics of written sentences (see e.g. Blaauw 1994, Schafer et al. 2000, Schegloff 1996, Shattuck-Hufnagel and Turk 1996). Thus, sentences, when spoken, can be chunked into pieces of information in variable ways. As such, they clearly pose a challenge for listeners whose comprehension processes are racing against the speed with which the acoustic details decay in working memory (Christiansen and Chater 2016). Therefore, the question is what listeners hear when processing spoken language and how they cope with strict time constraints on speech comprehension.

When processing spoken language, listeners are known to reduce the processing load of available acoustic information through the general process of chunking (Carpenter and Just 1989, Christiansen

* **Corresponding author: Nele Ots,** Institute of Linguistics, Goethe University, 60629, Frankfurt am Main, Hessen, Germany, e-mail: ots@em.uni-frankfurt.de
Piia Taremaa: Institute of Estonian and General linguistics, University of Tartu, 51005, Tartu, Estonia, e-mail: piia.taremaa@ut.ee

and Chater 2016). Speech chunking entails that smaller units (e.g. vowels, consonants, syllables) are rapidly engaged into and interpreted in larger ones to support the working memory (Christiansen and Chater 2016, McCauley and Christiansen 2015). Few studies of written language demonstrate evidence that perceptual language chunks correspond with syntactically defined elements, such as syntactic clauses or syntactic components (Grosjean et al. 1979, Niu and Osborne 2019). While Grosjean et al. (1979) have also proposed that perceptual language chunks may be defined by phonological factors like the length or weight of phrases, few studies have investigated to what degree perceptual language chunks might correspond with units of prosodic coherence or intonational phrases. For instance, Cole et al. (2010) investigated whether speech chunking is influenced by the distribution of prosodic information or by clausal structure. They found that even when acoustic cues are available (participants were able to listen to speech excerpts), the speech chunks identified by listeners naïve to the phonetic study of language largely correspond with the syntactic organization of spoken sentences. Cole et al. (2010) argued that the reason might be that prosodic variation in speech production tends to be largely modulated by the syntactic organization of sentences (see the numerous studies discussed in Cutler et al. 1997, Wagner and Watson 2010).

Nevertheless, relatively little is known about the interplay between prosodic and nonprosodic information in the process of perceptual speech chunking. The overarching goal of this study is to investigate to what degree the perception of syntactic clauses as speech chunks might be modulated by distributions of prosodic information that, at least partially, map onto syntactically defined segments of language. To achieve this goal, we first examine to what degree the syntactic clauses in conversational speech correspond with units of prosodic coherence. In particular, we aim to clarify the degree to which the distribution of prosodic information corresponds with the distribution of clause boundaries or, possibly, some other types of regular units (e.g. collocations) present in spoken sentences. Second, we investigate what listeners hear in spontaneous speech. More specifically, we aim to establish the impact of syntactic and prosodic information on perceptual chunking of conversational speech excerpts. A deeper understanding of the nature of speech chunks contributes to accounts of the cognitive mechanisms underlying the comprehension of spontaneous speech and language learning.

1.1 The production of prosody: structures underlying the prosodic units

Modelling speech comprehension processes requires understanding speech production because listeners, when apprehending spoken language, might integrate their knowledge of linguistic and non-linguistic constraints on language production (e.g. Clifton et al. 2006, Dahan 2015, Dahan and Ferreira 2019). With regard to speech production, messages, when verbally delivered, only seldom constitute sentences as they appear in written texts. More frequently, they are delivered in chains of speech chunks that can be recognized from systematic variations of pausing (intervals of silence), duration (perceived as speech rate and rhythm), intensity (perceived as loudness), and fundamental frequency (F0; perceived as pitch), which are collectively termed speech prosody. One well-known and intuitive cue for discerning prosodic speech chunks is intervals of silence or pauses (Cooper and Paccia-Cooper 1980, Grosjean et al. 1979, Krivokapić 2007, Strangert 1997). The chunks may be further characterized by lengthening and stronger articulation of segments at the beginnings of chunk-initial words (e.g. initial strengthening; Cho and Keating 2001, Keating et al. 2003, Oller 1973) and lengthening at the ends of chunk-final words. The latter phenomenon is frequently called pre-boundary lengthening, phrase-final lengthening, or just final lengthening (Berkovits 1994, Cambier-Langeveld 1997, Fon et al. 2011, Oller 1973, Turk and Shattuck-Hufnagel 2007, Wightman et al. 1992). An important tonal feature of a speech chunk is a continuous decline of F0 accompanied by intensity drop at the ends of chunks (Cooper and Sorensen 1981, Ladd 1988, Peters 1999, Thorsen 1985, Trouvain et al. 1998, Ulbrich 2002, Wagner and McAuliffe 2019). Moreover, the ends of chunks (the chunk-final words) may be characterized by upstepping pitch (Kentner and Féry 2013) or a rising boundary tone (O'Shaughnessy 1979, Petrone et al. 2017), both of which are indexed by high F0 at the final boundaries of prosodic chunks. Thus, relative to the centre of a prosodic speech chunk, the boundaries of a chunk are

characterized by the disruption of an established temporal and melodic structure. As such, prosodic speech chunks constitute units of prosodic coherence defined by clear prosodic discontinuities at their boundaries.

A central question in the study of sentence prosody has been to what degree the distribution of prosodic information may be conditioned by the syntactic organization of spoken sentences (see e.g. Berkovits 1994, Fon et al. 2011, Nakai et al. 2009, Wightman et al. 1992, for comprehensive reviews, see Cutler 1976, Wagner and Watson 2010). For example, several studies report that in subordinated or embedded clauses, F0 declination starts from a higher level than predicted from the F0 declination of the main clause (see e.g. Cooper and Sorensen 1981, Féry and Ishihara 2009, O'Shaughnessy 1979), indicating that clause boundaries trigger a pitch reset. In a study of major syntactic boundaries, Klatt (1975) reported pausing and pre-boundary lengthening at the boundaries between the subject noun phrases and predicates. Lehiste (1972), however, could not attest pre-boundary lengthening when comparing the lengths of syllables at the ends of subject noun phrases with the lengths of syllables preceding derivative suffixes (e.g. "the stick fell" vs "sticking"). Nevertheless, prosodic discontinuities occur at major boundaries where a constituent is locally ambiguous between the interpretations of a goal (e.g. "Put the dog | in the basket on the star") and a modifier (e.g. "Put the dog in the basket | on the star") (Kraljic and Brennan 2005, Lehiste 1973, Schafer et al. 2000, 2005, Snedeker and Trueswell 2003). Similarly, speakers are known to differentiate between two renditions of the type of the coordinating sentence "Steve or Sam and Bob will come" by producing a prosodic discontinuity after "Sam" or after "or" (Kentner and Féry 2013, Lehiste 1973, O'Shaughnessy 1979, Petrone et al. 2017, Turk and Shattuck-Hufnagel 2007). These findings suggest factors of the distribution of prosodic variation that are independent of the syntactic organization of spoken sentences.

Another type of abstract structure that has been widely accepted to exert control over prosodic variation in speech is the prosodic-metric structure (Beckman 1986, Beckman and Edwards 1990, Ladd 2008, Nespor and Vogel 1986, Selkirk 1984). The prosodic-metric structure is a type of abstract prosodic structure that has been proposed to mediate between prosodic information and syntactic structure and, importantly, to control for the distribution of prosodic information through a hierarchical structure of intonation phrases, phonological or accentual phrases, prosodic words, metrical feet, and syllables. In doing so, it nevertheless mainly refers to the syntactic structure (Nespor and Vogel 1986, Selkirk 1984, Truckenbrodt 1999; for an overview, see Dahan 2015). This especially holds true for the higher levels of the hierarchy. For example, clauses, that is verbs together with their arguments, constitute intonation phrases. Intonation phrases divide into smaller intonational units that mostly align with major syntactic phrases (e.g. subject noun phrases). In other words, the theory of prosodic-metrical structure proposes that the prosodic organization of the supralexical levels (i.e. phonological, accentual, and intonational phrases) in spoken sentences to a large degree overlaps with the syntactic phrase structure. Few studies have made efforts to establish independent phonological or prosodic-metric factors (Beckman and Edwards 1990, Cambier-Langeveld 1997, Grosjean et al. 1979, Keating 2006, Turk and Shattuck-Hufnagel 2007, Watson and Gibson 2005), but the evidence that prosodic structure controls the distributions of prosodic information independently from syntactic structure is still notably scarce. This research goal, though, remains outside of the scope of the current study due to the limits of our analysis, which did not consider abstract categories of pitch accents and boundary tones.

One of the explanations for the relatively tight relationship between syntactic organization and prosodic structure is that speakers may be aware of their addressees and may wish to ease language comprehension processes by highlighting the intended syntactic structures (see e.g. Berkovits 1994, Fon et al. 2011, Hawthorne 2018, Lehiste 1973, Nakai et al. 2009, Petrone et al. 2017, Turk and Shattuck-Hufnagel 2007, Wightman et al. 1992). Investigations, however, indicate that the audience has a weak or even no effect in the production of prosodic discontinuities (Kraljic and Brennan 2005, Schafer et al. 2000, Snedeker and Trueswell 2003). In particular, multiple tests by Kraljic and Brennan (2005) revealed no effect from the presence of a listener, a communicative goal, or syntactico-semantic ambiguity in the message to be conveyed. In particular, speakers in their studies produced disambiguating prosody (pauses and/or pre-boundary lengthening) independently of whether the picture they were asked to describe showed an ambiguous scene, whether they were aware of the addressee's perspective (when performing an action in the second block of the experiment after being instructors themselves), and whether they were speaking

alone or to a partner. Crucially, these results led Kraljic and Brennan (2005) to a conclusion that the production of prosodic discontinuities at syntactic phrase boundaries in a spontaneous speaking task arises from "planning and articulating the syntactic structure" rather than from consideration of the addressees' needs.

Accordingly, several accounts of language production share the related theoretical view that speakers might plan their utterances in stages that correspond with the levels of linguistic analysis, i.e. the conceptualization of the messages is followed by the generation of syntactic representation, which, in turn, guides lexical and phonological processing (Bock et al. 2004, Bock and Levelt 1994, Levelt 1989). In particular, Wheeldon and Smith (2003), and Wheeldon et al. (2013) have convincingly demonstrated that the unit of planning spans words that constitute a syntactic phrase (e.g. a subject noun phrase). Moreover, intonation contours are often preserved despite slips of the tongue, and regular assimilations and dissimilations of speech sounds (e.g. a change from /p/ to /m/) do not cross syntactic phrase boundaries (Fromkin 1971, Keating and Shattuck-Hufnagel 2002). This has been taken to indicate that prosodic structure is planned based on the syntactic structure and before phonological encoding (Fromkin 1971, Keating and Shattuck-Hufnagel 2002). Thus, the distribution of prosodic information in spoken sentences might indeed reflect syntax-driven production processes (for the speech prosody to index the cognitive architecture of language production, see also Bürki (2018)), which is an idea that was already present in early investigations of pre-boundary lengthening (see e.g. Oller 1973).

1.2 Perception of prosody: what do listeners hear?

Listeners are greatly affected by prosodic discontinuities. For example, several recent studies measuring brain signals have detected significant brain activity time-locked to the locations of prosodic discontinuities (Bögels et al. 2010, 2011, 2013, Kerkhofs et al. 2008, Steinhauer et al. 1999). In particular, Kerkhofs et al. (2008) compared brain activity between reading the conventional placement of commas and listening to spoken sentences controlled for prosody. They found that compared to reading commas, only prosodic discontinuities evoked the attentional component closure positive shift of the ERP (event-related potential) component in the brain. Moreover, an unexpected prosodic discontinuity may even impede speech comprehension processes (Bögels et al. 2013).

Perceptual study of prosodic information has confirmed that prosodic discontinuities in production constitute prosodic breaks (i.e. prosodic phrase boundaries) in perception. Several studies have explicitly asked listeners to indicate a break or disjunction between two words (Petrone et al. 2017, de Pijper and Sanderman 1994, Streeter 1978, Yang et al. 2014). The results demonstrate that, for listeners who are naïve to making any explicit reference to phonetic events, a disjuncture is likely to be reported at the locations of pauses, longer syllable rhymes, pitch resets, and rising F0 movements (see e.g. de Pijper and Sanderman 1994). In other words, non-expert listeners indeed report hearing a break at places of prosodic discontinuities. Although some accounts pose the important question of whether listeners might attend to prosodic coherence rather than prosodic breaks (e.g. Schafer 1997, Watson and Gibson 2005; for discussion, Wagner and Watson 2010), phonetic perception experiments establish that pausing, a relative discontinuity in segment duration, intensity, and pitch curve indeed underlie the significant percept of a prosodic phrase boundary or a prosodic break (Petrone et al. 2017, de Pijper and Sanderman 1994, Streeter 1978, Yang et al. 2014).

Furthermore, different types of prosodic discontinuities contribute to the percept of a prosodic break to different degrees. While pausing most directly contributes to the perception of a prosodic break (Himmelmann et al. 2018, Petrone et al. 2017, Riesberg et al. 2020, Simon and Christodoulides 2016, Yang et al. 2014), durational and melodic discontinuities appear to have a weaker influence (Männel and Friederici 2016, Peters et al. 2005, Yang et al. 2014). In some studies, pre-boundary lengthening, intensity and F0 curves appear to contribute to the perception of prosodic boundaries only in combination (Männel and Friederici 2016, Peters et al. 2005, Yang et al. 2014), whereas in other studies, durational and tonal

variations cue prosodic breaks independently of pauses and each other (Peters 2005, Streeter 1978, Swerts et al. 1994). More recently, Petrone et al. (2017) investigated the individual contributions of melodic and durational discontinuities by rigorously controlling for each of the cues (pausing, duration, pitch). They demonstrate that each of these cues has an independent influence on prosodic boundary perception. In addition, they show that the duration of a pause is interpreted more categorically than the duration of segments and changes in the pitch range. Petrone et al. (2017) used these results to indicate that pauses have a stronger effect than pre-boundary lengthening and F0 curves on the perception of prosodic breaks. In addition, Hawthorne (2018) showed that flat F0 contours (obtained with the method of noise vocoding) did not impede the recognition and memory of novel words, which suggests in support of Petrone et al. (2017) that pausing and pre-boundary lengthening still present in the noise-vocoded speech may play a greater role than melodic discontinuities in the auditive processing of prosodic breaks.

In linguistic processing, prosodic phrase boundaries have been shown to support the decoding processes and possibly to speed up speech comprehension. For example, words in a novel/artificial language are learned better when they co-occur with pauses or pre-boundary lengthening, indicating that prosodic discontinuities (prosodic phrase boundaries) might help with rapid recognition of lexical segments of continuous speech flow (see e.g. Christophe et al. 2004, McQueen and Cho 2003, White et al. 2020). Recent studies demonstrate the role of prosodic phrase boundaries in processing at higher linguistic levels as well (Hawthorne 2018, Langus et al. 2012, Ordin et al. 2017). For example, utilizing the paradigm of learning an artificial language, Langus et al. (2012) demonstrated that continuous downdrift of F0 across a segment of speech and longer segment durations induce the recognition of words that somehow belong together, i.e. form a semantic and/or syntactic unit. These results indicate that the prosodic phrase boundaries in spoken language not only facilitate learning the semantic and syntactic relationships between the words (as concluded in Langus et al. 2012), but also may enable rapid access to the semantic and syntactic organization of spoken sentences in the process of natural speech decoding.

Indeed, prosodic phrase boundaries are very useful for assessing the intended meanings of spoken sentences (e.g. Kraljic and Brennan 2005, Price et al. 1991, Schafer et al. 2000, Snedeker and Trueswell 2003). For example, participants in a study by Price et al. (1991) were asked to listen to sentences spoken by four professional public radio broadcasters and to choose between two contexts from which these sentences might have originated. The results of this forced-choice task indicate that listeners may recover the original context of speech based on prosodic information. Namely, acoustic analysis of the constituent boundaries critical to the meanings of the sentences showed that the two different readings were mostly characterized by pre-boundary lengthening and pausing. In another type of study, participants performed an action upon attending to instructions delivered by a speaker (e.g. Kraljic and Brennan 2005, Schafer et al. 2000, Snedeker and Trueswell 2003). Listeners were required to correctly interpret instructions in syntactically and semantically ambiguous sentences such as “Tap the frog with a flower.” While the instructions in the study by Snedeker and Trueswell (2003) were spoken by a professional speaker, they were spoken by other participants of the study by Kraljic and Brennan (2005), and Schafer et al. (2000). Regardless of whether the instructions came from the professional speaker, listeners in these studies correctly interpreted the presence of a prosodic boundary after the first noun phrase (e.g. “frog”), especially in the study by Kraljic and Brennan (2005). Thus, discerning the correct location of prosodic breaks may be vital for the decoding the meaning of verbal messages.

In an attempt to establish the independent cognitive function of prosodic information, several studies have discovered that listeners’ sensitivity to prosodic breaks is strongly modulated by the syntactic structure of spoken sentences (Duez 1985, Simon and Christodoulides 2016). For instance, Simon and Christodoulides (2016) asked listeners to press a button as soon as they hear a break or some sort of a juncture while listening to normal or delexicalized audio recordings of spontaneous speech. After aligning the button presses with the corresponding speech signals, Simon and Christodoulides (2016) found that more prosodic breaks were reported for normal speech than for delexicalized speech. Therefore, even when they are explicitly asked to pay attention to acoustic information, listeners appear to rely on the semantic and syntactic organization of spoken sentences. In a more controlled setting, Buxó-Lugo and Watson (2016) demonstrated that prosodic discontinuities that coincide with syntactic clause boundaries

contribute more strongly to the perception of prosodic breaks than those occurring within clauses. In addition, listeners in a study by Cole et al. (2010) were explicitly asked to chunk excerpts of spontaneous speech. Their task was to listen to spontaneously spoken speech excerpts and to indicate any breaks or junctures in the transcripts of these excerpts. The results show that speakers were more likely to report a break after a word that was at the clause boundary. While the word durations were longer in the presence of breaks than in the absence of breaks, the F0 maxima did not vary as a function of boundary perception. These results warrant the conclusion that the perception of clause boundaries as prosodic breaks might be modulated by pre-boundary lengthening, as pre-boundary lengthening is highly likely to occur at clause boundaries (Cole et al. 2010). It appears thus that similar to production of prosody, the perception of prosody is also tightly integrated with syntactic information.

1.3 Towards a model of online speech comprehension

A comprehensive model of speech perception is expected to be anchored, at least partly, in processes of speech production (see e.g. Dahan 2015). It appears that the production of prosody frequently reflects the syntactic organization of spoken sentences, but this relationship between the distribution of prosodic information and syntactic structure is by no means compulsory. Plentiful evidence suggests that in spoken sentences or utterances, the prosodic phrase boundaries are less likely to concur with syntactic constituent boundaries than in read-aloud sentences (see e.g. Blaauw 1994, Schafer et al. 2000, Schegloff 1996, Shattuck-Hufnagel and Turk 1996). Prosodic discontinuities often arise from difficulties in production processes, and in such cases, they are not expected to be semantically or syntactically motivated (Ferreira and Karimi 2015). Moreover, speakers appear not to be aware of their audience even while producing prosodic breaks that turn out to be helpful for listeners (see e.g. Kraljic and Brennan 2005). Thus, speakers' production of prosodic breaks may depend on the availability of cognitive resources and the degree to which the syntactic structure might have guided the production processes (e.g. Bock et al. 2004, Konopka and Meyer 2014, Levelt 1989) rather than on speakers' awareness of listeners' needs. Despite the high variability in prosody–syntax mapping, the semantic and syntactic processing of spontaneously spoken sentences appears to greatly benefit from the distributions of prosodic information (see the previous section). Thus, effective utilization of distributions of prosodic information may be based on listeners' inferences of physiological as well as linguistic constraints on the speakers' production processes (e.g. Clifton et al. 2006, Dahan 2015, Dahan and Ferreira 2019).

Speech comprehension can be viewed as a task of "breaking continuous streams of sounds into units that can be recognized" (Sanders and Neville 2000, p. 1) and maintained in the working memory for further processing (Christiansen and Chater 2016). In other words, for speech comprehension processes, the continuous flow of speech needs to be broken into cognitive units, that is chunks of language. To be effective, language chunking needs to rely not only on world knowledge about causalities and experience with communicative situations but also on linguistic knowledge. Within the chunking process, access to and resolution of linguistic knowledge has been proposed to reflect predictive modelling procedures (Clark 2003, Denham and Winkler 2006). More specifically, Dahan and Ferreira (2019) posit that based on a given sensory input, listeners generate a set of hypotheses about the prosodic, syntactic, and semantic structures that might have generated a particular speech signal. Most likely, such hypotheses are generated and evaluated based on speaker-internal production processes, as nearly every listener is also a speaker (Clifton et al. 2006, Dahan 2015). By constantly re-evaluating and narrowing down their hypotheses, listeners incrementally recover the intended prosodic, syntactic, and semantic organization of spontaneously produced utterances, and they also usually arrive at meanings shared with the speaker.

Relying on this model, we propose that prosodic information is vital to the auditory segmentation of spoken language by shaping the sensory input that listeners use to generate the hypotheses about linguistic structures that potentially correspond to a given speech signal. The activation of these linguistic structures enables listeners to effectively ignore prosodic information that does not concur with the syntactic and

semantic organization of utterances but that might arise for other reasons (e.g. due to difficulties in planning processes; Ferreira and Karimi 2015). Thus, the prediction that we draw from the model in Dahan and Ferreira (2019) is that prosodic information, such as acoustic pauses, rhythmic accelerations, decelerations, and abrupt intonational upsteps, remains unnoticed by listeners when it does not coincide with some sort of structural unit (e.g. an abstract prosodic category, syntactic clause, or semantic component) because they are rejected as cues to linguistically relevant structures as a result of rapid and internal hypothesis testing.

1.4 The current study

The goals of the current study are two-fold. First, we aim to examine based on the excerpts excised from spontaneous dialogues, to what degree the distribution of prosodic information is influenced by syntactically defined (i.e. syntactic clause and phrase boundaries) and probabilistically determined regular segments of language (i.e. frequent sequences of two or three words, bigrams and trigrams, respectively). Most of the earlier studies reported in previous sections investigate read-aloud speech of trained or untrained speakers (e.g. O'Shaughnessy 1979, Petrone et al. 2017, Price et al. 1991, Wightman et al. 1992) or spontaneous speech elicited under strict conditions of predefined sentence constructions or highly controlled communicative goals (Kraljic and Brennan 2005, Schafer et al. 2000, 2005, Snedeker and Trueswell 2003). In contrast, we investigate excerpts of naturally spoken utterances from a corpus of spontaneous dialogues spoken in Estonian (Lippus et al. 2016). The selected excerpts reflect the natural speech conversations to a highest degree. Namely, two speakers, untrained in public speaking, who are friends or good acquaintances chat on a freely chosen topic, while they are recorded in a sound-attenuated professional recording studio. The choice of our materials enables us, thus, to examine whether the findings established based on the laboratory speech generalize on to the more natural speech conditions.

Second, we investigate to what degree the perception of prosodic breaks corresponds with the abstract syntactic structure or with signal-based prosodic information (prosodic discontinuities) based on the excerpts of natural conversations. We expect that the syntactic organization of utterances strongly modulates the perception of prosodic discontinuities as prosodic breaks. The analysis of speech production will enable us to determine whether the strong impact of syntactic structure is caused by the prosodic discontinuities regularly aligning with syntactic constituent boundaries (like proposed in Cole et al. 2010). To find support for the proposed model of speech comprehension (see Section 1.3), the correspondence between the distribution of prosodic information and syntactic structure in production should be rather weak. We expect the effect of syntactic structure despite the highly variable syntax-prosody mapping because according to the model, the prosodic discontinuities (i.e. signal-based information) should decay fast in the working memory if they turn out as no cues to the linguistic structures.

To further exhaust the role of syntactic representation in the speech comprehension processes, we compare its impact with the influences from frequent sequences of two or three words, the so-called lexical collocations. Language production, especially articulation, and comprehension are necessarily sequential processes. It may be that elements that are frequently used together evolve into some types of constituents (Bybee 2002) that might but do not need to coincide with syntactically defined segments of language. For instance, a recent computer-linguistic study has applied the collocation frequencies with some success in the task of automatic and unsupervised segmentation of large text corpora (see Borrelli et al. 2020). This finding may be taken to indicate that collocations constitute indeed some sort of a unit or even a constituent. More generally, the frequent lexical collocations reflect on the natural language usage and they can be taken as indexes of users' exposure to closely related words. As such, they may, similarly to syntactic constituents, modulate the variation of prosodic information but also be useful for the language comprehension processes.

In the following, we present two studies investigating speech chunking in relation to syntactic constituent structure and collocation frequencies: a corpus study and a perception study. The corpus study,

like a number of previous studies, investigates the acoustics of the constituent boundaries in comparison to non-boundaries. Furthermore, it examines whether the increasing likelihood of a collocation influences durational, tonal, and intensity variation. The perception study tests to what degree prosodic information, syntactic structure, and collocation frequencies influence the perception of prosodic breaks, or more generally, perceptual speech chunking. With this, the aim is to tap into a few of the linguistic abstractions on which the comprehension of continuous speech flow may rely on.

2 Corpus study

2.1 Methods

2.1.1 Materials

A stretch of fluent speech between silent intervals of 400 ms or longer was defined as an utterance. A number of spontaneously spoken utterances (N = 396) from randomly selected ten speakers (five females, average age 25.3 years) were drawn from the phonetic corpus of spoken Estonian (Lippus et al. 2016). The selection of utterances for the analysis was based on a set of arbitrary phonetic, phonological, and performance criteria. First, the number of syllables was allowed to vary between 18 and 24. Second, utterances containing many or long stretches of disfluencies were excluded. The average duration of utterances included in the analysis and further experimentation (perception study, see Chapter 3) was 3,300 ms.

2.1.2 Acoustic analyses

The analyses involved four different dependent variables. First, for each word, we automatically extracted the duration of the last syllable (in milliseconds; *Syllable Duration*). The absolute duration of the last syllable is taken to index the pre-boundary lengthening.

Second, since the utterance was defined to be a stretch of fluent speech between the silent intervals of 400 ms or longer, the selected utterances still contained some pauses and some minor hesitations shorter than 400 ms. The duration of these silent and filled pauses was automatically extracted and included in the analysis (in milliseconds; *Pause Duration*).

Third, F0 (in hertz) was extracted from utterances with the help of the auto-correlation method available in Praat (Boersma and Weenink 2020) in two passes. During the first pass, F0 contours were extracted with default settings for the lowest and highest F0, the "floor" and "ceiling" (75 and 600 Hz, respectively). Then, the first and third quartiles of F0 (Q1 and Q3) were calculated for each speaker and recorded in a table. In the second pass, F0 contours were extracted with speaker-specific settings (0.75*Q1 for floor and 1.5*Q3 for ceiling). For the analysis, the F0 contours were converted into semitones by using the equation (1)

$$F0_{\mathrm{st}} = 12 * \log_2\left(\frac{F0_{\mathrm{Hz}}}{F0_{\mathrm{mean}}}\right), \tag{1}$$

where $F0_{\mathrm{mean}}$ is the speaker's mean F0 aggregated over all F0 contours extracted from the speaker's utterances.

To approximate pitch reset (in Hz; *Declination Estimate*), for each utterance, a regression line was fitted to F0 contour as a function of time-called as utterance declination. Another set of domain-specific declinations was determined by fitting regression lines to F0 contours of the domains of interest, i.e. clauses, phrases, bigrams, trigrams (corpus study), and chunks detected by the participants of the experiment (perception study; for the methodology of detecting melodic discontinuities by fitting regression lines, see e.g. Beňuš et al. 2014, Reichel 2011). The pitch level at the beginning of each domain (i.e. a clause, a

phrase, a bigram, a trigram, and a perceptual chunk) was F0 (in semitones) as estimated by the domain-specific regression lines. In the analysis, the domain-specific pitch level is compared against the pitch level as predicted by the utterance declination. If the declination trend is reset at the beginning of a relevant domain, then the *Declination Estimate* of this domain should be higher than the *Declination Estimate* estimated from the utterance declination.

Fourth, we investigated F0 at the boundaries of the relevant domains of interest by calculating relative F0 peaks (in Hz; *Relative F0*). Namely, a number of studies have detected that the F0 level in the words at the boundaries of clauses is high (Kentner and Féry 2013, Petrone et al. 2017). While Petrone et al. (2017) measured the F0 maxima at the word-final syllables located at the clause boundaries, Kentner and Féry (2013) measured a mean F0 of words at the boundary. Both analyses arrive at the result indicating that the utterance-internal boundary of a clause is accompanied by high F0 maxima indexing final rise. To approximate the F0 at the boundaries of the relevant domains, we automatically extracted F0 maxima from the pre-boundary words. To derive a relative F0 measure, the mean F0 was divided with the average F0 of corresponding utterances. The positive or high values of *Relative F0* are taken to index tonal rises. For the tonal rise or upstep to be present, *Relative F0* should be higher at the tonal boundaries than at no boundaries.

Finally, the intensity as root mean square (RMS) amplitude of the very first and the very last syllable of a word was automatically extracted. For the approximation of the intensity curve within a word, intensity difference was obtained by subtracting the RMS value of the last syllable from the RMS value of the first syllable (*Intensity Difference*). The larger the intensity difference between the first and the final syllable, the larger is the intensity drop across the word.

2.1.3 Syntactic scoring

In general, spontaneously spoken utterances pose a great challenge for syntactic analyses, as they often contain uncanonical placements of words, unfinished and elliptical utterances, parenthesis, mispronunciations, truncated words, hesitated and filled pauses, repetitions, false starts, self-repairs, and other phenomena specific to spoken language (Müürisep and Nigol 2008). In addition, the word order is relatively free in Estonian (Lindström 2006, Tael 1988, Vilkuna 1995), such that it is difficult to detect a verb phrase that is not intervened by some verb phrase modifier, focus particle, or parenthetical. Therefore, the syntactic scoring was not devised in a particular syntactic framework but followed the functional framework in the widely accepted description of Estonian syntax (Erelt and Metslang 2017).

This analysis concentrates on the syntactic function of noun phrases and simply determines the location of the finite verb instead of the verb phrase because verbs are frequently composed of multiple parts that can easily be separated and located anywhere in sentences/utterances. Therefore, our analysis was limited to the determination of clauses and major syntactic phrases. A clause was defined to be the finite verb together with its obligatory and/or optional argument and was allowed to encompass also non-constituents such as disclosures and interjections. For the syntactic phrases, we first categorized the noun phrases as belonging to six types of syntactic functions such as subject, object, adverbial, predicative, and predicative adverbial (see column “Function” in Table 1), while following the syntactic and semantic criteria suggested in the account of Estonian syntax by Erelt and Metslang (2017). A noun phrase was left uncategorized when it was not possible to determine its syntactic function due to a repair, ellipsis, false start, or truncation. Then, based on this scoring of noun phrases, we approximated the boundaries of the subject noun phrases and the verb phrases. The verb phrase was determined to include and/or span over the finite verb, the infinite verb, the object noun phrase, and the predicate. For the syntactic structure, always the last words of corresponding units (a phrase or a clause) were tagged as boundaries of these units (Table 1).

For collocation frequencies, all words within excerpts were combined with either a single word or two words preceding it. The last word of the pair or triplet was associated with a measure of collocation frequency based on the corpus of Estonian written language (Raudvere and Uiboaed 2018). In general,

Table 1: Sample scoring of an utterance *ja siis käisime seal Iisraeli muuseumis, kus see suur makett oli, mis oli päris võimas* "Then we went to this Israelian museum, where this big maquette was, which was pretty awesome."

Row	Transcription	Translation	Function	Clause boundary	Phrase boundary
1	*ja*	and	conjunction	0	0
2	*siis*	then	adverbial	0	0
3	*käisi-me*	went-we	verb	0	1
4	*seal*	there	adverbial	0	0
5	*iisraeli*	Israelian	adverbial	0	0
6	*muuseumi-s*	museum-in	adverbial	1	0
7	*kus*	where	conjunction	0	0
8	*see*	this	subject	0	0
9	*suur*	big	subject	0	0
10	*makett*	maquette	subject	0	1
11	*oli*	was	verb	1	1
12	*mis*	which	conjunction	0	0
13	*oli*	was	verb	0	0
14	*päris*	pretty	predicative	0	0
15	*võimas*	awesome	predicative	1	1

The absence of boundaries is indicated by a zero sign and the presence of boundaries is noted with a symbol of 1. The phrase boundary refers to the boundaries of subject noun phrases and verb phrases.

the collocation frequencies found in our data were pretty low. The increasing likelihood is interpreted as a likely boundary of a collocation. That is, the higher the measure of collocation likelihood of a corresponding word, the likelier is the boundary of a collocation in the given word.

2.1.4 Evaluation

The statistical analysis aimed to determine how likely is the boundary of a relevant domain (i.e. a clause, a phrase, a bigram, or a trigram) given the distribution of a particular type of prosodic information (i.e. the duration of pauses, the duration of word-final syllables, changes in pitch and intensity curve). Therefore, the four mixed effects regression models were defined to estimate to what degree the five continuous variables – *Pause Duration*, duration of word-final syllable (*Syllable Duration*), *Declination Estimate*, *Relative F0* of the word-final syllable, and *Intensity Difference* – contribute to the production of clause boundaries, phrase boundaries, bigram boundaries, and trigram boundaries.

Clause and phrase boundaries constitute respective dependent variables with binomial distributions, because a boundary can either be present or not (scored with 0 and 1, respectively, Table 1). Therefore, we defined two general linear mixed effects regression (GLMER) analyses separately for the dependent variables of clause and phrase boundaries. The analyses estimated whether the predictor variables, referred to as explanatory variables *Pause Duration*, *Syllable Duration*, *Declination Estimate*, *Relative F0*, and *Intensity Difference*, contribute to the likelihood of a syntactic boundary. The duration of word-final syllables was logarithmically transformed with the base of 10. There was no need to transform the duration of pauses because the distribution of *Pause Duration* resembled the normal distribution quite well. All the continuous explanatory variables (*Pause Duration*, *Syllable Duration*, *Declination Estimate*, *Relative F0*, *Intensity Difference*) were z-scored, that is scaled and centred for the regression analysis. The scaling and centring procedures enable us to compare between the effect sizes of the five different explanatory variables.

The boundaries of two-word sequences and three-word sequences, called as bigrams and trigrams respectively, are indexed by the frequencies of occurrence in the corpus of fictional texts (Raudvere and Uiboaed 2018). The greater the frequency, the likelier the boundary. As such, the frequencies of bigrams and trigrams constitute the continuous dependent variables that allow us to test the linear relationships between the collocation boundaries and the prosodic discontinuities. Therefore, we determined two linear

mixed effects regression (LMER) analyses separately for the dependent variables of bigram and trigram frequencies. Similar to the duration of pauses and syllables, frequency of the bigrams and trigrams was logarithmically transformed with the base of 10. Unfortunately, the logarithmic transformation of trigram frequency did not remove the left skew of the distribution, most probably, due to a great number of trigrams of zero frequency in the corpus in Raudvere and Uiboaed (2018). All the continuous explanatory variables (*Pause Duration*, *Syllable Duration*, *Declination Estimate*, *Relative F0*, *Intensity Difference*) were z-scored.

The analyses were conceptualized to estimate the size of the effect while holding the effects originating from other prosodic variable constants. In doing this, we opt to a type of multiple regression analysis, where the variables are not dropped for finding the best model and possibly the higher significance values (for the inadvisability of stepwise regression methods, see e.g. Winter 2019, 276–277). The analyses were run as implemented by the *lmer* package (Bates et al. 2015) available in the R software (Core Team 2019). In addition to test variables, all models included random slopes for speakers. The converging model fit was obtained by using the *optimx* optimizer (Nash 2014, Nash and Varadhan 2011). The significances of the model estimates were accomplished using the *jtools* package (Long 2017). The diagnostics of GLMER fits was conducted with the *DHARMa* package (Dunn and Smyth 1996, Gelman and Hill 2006).

2.2 Results

The five continuous variables (*Pause Duration*, *Syllable Duration*, *Declination Estimate*, *Relative F0*, *Intensity Difference*) were defined to predict the occurrence of a clause boundary, phrase boundary, bigram boundary, and trigram boundary. The correlations between the explanatory variables were estimated by calculating Pearson's r coefficients (Table 2). The correlations between the domain-specific *Declination Estimates* of clauses, phrases, bigrams, and trigrams were expectedly high. They are excluded from Table 2 because they function in the separate models and the correlations between them are not of particular interest.

Pearson's r coefficients in Table 2 indicate low correlations between the *Declination Estimates*, *Pause Duration*, *Syllable Duration*, *Intensity Difference*, and *Relative F0*, which enables us to simultaneously investigate them as factors of clause boundary, phrase boundary, bigram boundary, or trigram boundary.

The results of the GLMER and LMER analyses are illustrated in Figure 1. The GLMER tested how likely is the clause boundary (blue points and blue lines in Figure 1) or the phrase boundary (red squares and red lines in Figure 1), given the prosodic variation as indexed by duration of pauses and word-final syllables, by F0 declination, relative F0 peaks, and intensity difference. The likelihood of a boundary was expected to increase with the increasing *Pause Duration* and *Syllable Duration*, and with higher *Declination Estimate*, *Relative F0*, and larger *Intensity Difference*. LMER analyses diagnosed whether the greater frequency of the bigrams (green diamonds and green lines in Figure 1) and trigrams (purple triangles and purple lines in Figure 1) is related to the longer *Pause Duration* and *Syllable Duration*, and to higher *Declination Estimate*, *Relative F0*, and larger *Intensity Difference*. The estimates below zero indicate a decrease of the dependent

Table 2: Pearson's r correlations between the five explanatory variables

	Decl. Clause	Decl. Phrase	Decl. Bigram	Decl. Trigram	Pause Dur.	Syl. Dur.	Int. Dif.
Decl. Clause							
Decl. Phrase							
Decl. Bigram							
Decl. Trigram							
Pause Dur.	−0.14	−0.15	−0.06	−0.06			
Syl. Dur.	−0.05	0.06	0.08	0.08	0.02		
Int. Dif.	0.22	0.15	0.20	0.20	−0.24	−0.15	
Rel. F0	−0.32	−0.07	−0.16	−0.16	0.12	0.13	−0.03

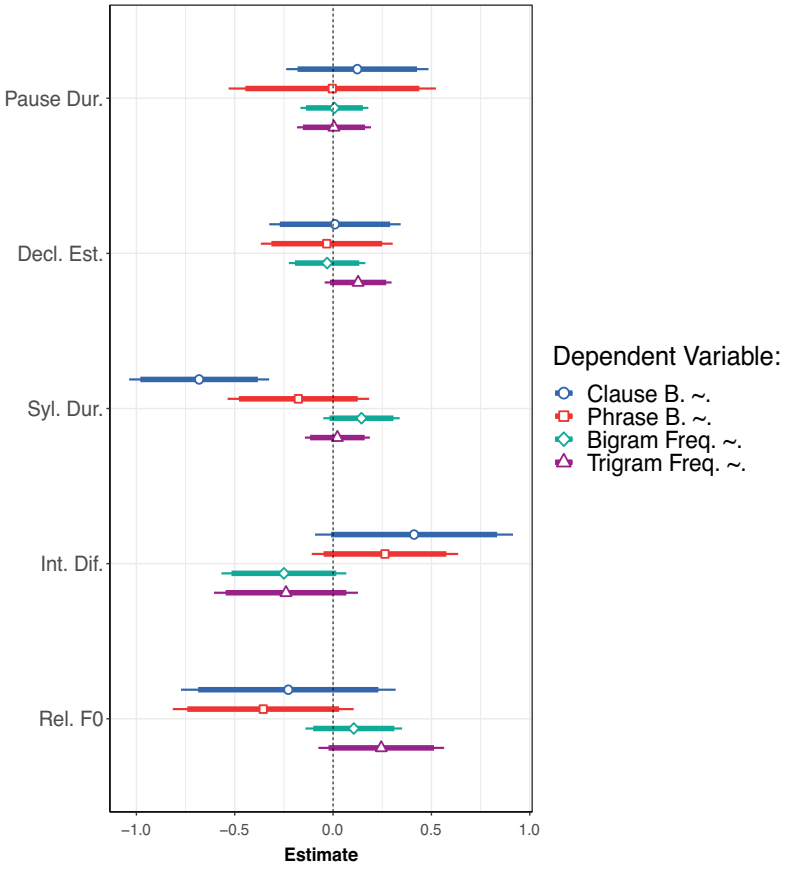

Figure 1: Contributions of *Pause Duration*, domain-specific *Declination Estimate*, word-final *Syllable Duration*, *Intensity Difference*, and *Relative F0* to the production of clause boundaries (blue points and blue lines), phrase boundaries (red squares and red lines), bigram boundaries (green diamonds and green lines), and trigram boundaries (purple triangles and purple lines). The points are the regression coefficients and the lines indicate the 95% confidence intervals of the coefficients. The further away is the regression coefficient (Estimate) from zero (vertical line), the larger is the effect.

variable as a function of an explanatory variable and the estimates above zero indicate an increase of the dependent variable as a function of an explanatory variable. For the positive effects of the explanatory variables and the support of our hypotheses, we wish to see values above zero (that is to the right of the vertical lines in Figure 1).

The GLMER analysis brought to light a rather weak correspondence between the distribution of different types of prosodic information and syntactic clause boundaries (the R^2 for the variance explained by the explanatory variables was 0.18; the R^2 for the entire model was 0.32). Figure 1 indicates that the clause boundary is more likely as the duration of the word-final syllable decreases (Est. = −0.68, CI = (−1.04, −0.33), $z = -3.75$, $p < 0.001$). The other variables (*Intensity Difference*, *Pause Duration*, *Declination Estimate*, and *Relative F0*) did not significantly contribute to the presence of clause boundaries.

The GLMER fit of phrase boundaries and explanatory variables indicates no relationship between the prosodic discontinuities and the likelihood of a phrase boundary (the R^2 for the variance explained by the explanatory variables was 0.05; the R^2 for the entire model was 0.20). None of the variables (*Intensity Difference*, *Pause Duration*, *Syllable Duration*, *Declination Estimate*, and *Relative F0*) contributed to the presence of a phrase boundary.

The LMER analysis of the relationship between the bigram frequency and the prosodic variables yielded no significant effects (the R^2 for the variance explained by the explanatory variables was 0.08; the R^2 for the entire model was 0.26). Similarly, LMER analysis did not detect significant contributions of prosodic discontinuities to trigram boundaries either (the R^2 for the variance explained by the explanatory variables was 0.15; the R^2 for the entire model was 0.80).

Table 3: Mean values of the word-final syllable duration (ms) split by the internal and final positions within utterances and clauses

	Utterance-internal	Utterance-final
Clause-internal	155.2	267.8
Clause-final	170.8	176.3

2.3 Interim discussion

The results indicate that much of the prosodic variation remains unexplained by the syntactic constituent structure and collocation boundaries. The only significant result concerns the clause boundaries. In particular, *Syllable Duration* turned out to significantly contribute to the production of clause boundaries. However, the effect of *Syllable Duration* was in an opposite direction of the prediction such that the shorter was the word-final syllable, the likelier was the clause boundary. For an explanation we consider that the utterance-final position interacts with the clause-final position in the determination of the pre-boundary syllable duration (Table 3).

In Table 3, we can observe that word-final syllables in the clause-internal positions and at the ends of utterances have the longest duration. Most likely, the greater average of the duration of clause-internal and utterance-final syllables affects the estimations of GLMER. Probably, these extremely long syllable durations reflect sustaining the "floor" of speaking or difficulties in speech planning. Unfortunately, the exclusion of the utterance-final words worsened the model fit. Thus, we refrain from this practice and leave the result as it is.

In sum, the spontaneous production of prosody appears to be conditioned by a number of factors. The spontaneous utterances contain much larger number of different types of non-constituents such as disclosures and interjections that might attract the occurrence of prosodic discontinuities more likely than the syntactic boundaries or the boundaries of the collocations. Moreover, difficulties in planning for speech production are known to cause pausing and considerable lengthening of the segments (see e.g. Ferreira and Karimi 2015, Lee et al. 2013). The remaining question is whether listeners naïve to prosodic study of language show fine-tuned sensitivity to all of prosodic variation or just the variation occurring at the syntactic boundaries and the collocation boundaries.

3 Study 2

The results of the production study indicate a considerable variation of prosody in spontaneously spoken utterances. In this section, we are interested in what guides the perception of prosodic breaks in the natural speech processing task. The aim of the experiment was to test the prediction that listeners are sensitive to prosodic discontinuities occurring at the structurally salient positions, that is at the boundaries of clauses, phrases, and collocations. The perception of prosodic breaks was tested in two modes of language processing. The first mode, called listening mode, involved listening to spontaneous utterances and rating the sequences of word pairs for disjunctures based on the transcriptions of these utterances. The second mode, called reading mode, involved reading and rating the word pairs based on the transcriptions of same utterances, without an access to acoustic information. The aim of the mode manipulation is to determine, to what degree the perception of syntactic boundaries as prosodic breaks may be modulated by the prosodic discontinuities.

3.1 Methods

3.1.1 Materials

The materials described in Section 2.1.1 served as the stimuli for the perception experiment. A list of 396 utterances was randomly distributed between the four lists of experimental items containing 99 utterances

each. In total, 4,726 words were rated by the different numbers of listeners and readers. We excluded the utterance-final words (N = 396) from further analysis because we did not explicitly ask participants to mark boundaries at the ends of utterances.

3.1.2 Participants

The experiment involved two groups of participants. The first group of participants included 47 native speakers of Estonian (38 females), average age 30 years (between 19 and 54). The contribution of the first group was voluntary. Since the experimental software employed does not enable controlling the assignment of participants to different lists of experimental items, the four lists of utterances were listened to by the different numbers of participants (the minimum number of listeners per list was 9 and the maximum number was 14). The second group of participants was recruited with the help of a crowd-sourcing marketplace designed for conducting research (Prolific). They were paid £2.13 for the completion of the study. In total, 40 native speakers of Estonian participated in the study (26 females), average age 27 years (between 18 and 60 years). For the assignment of the lists, the equal number of listeners per list (10) was achieved by incrementally increasing the study places in Prolific and changing the parameters of the experiment for the next 10 participants. All participants originated from different regions of Estonia. They did not report hearing or vision impairments.

3.1.3 Procedure

The experiment was conducted over the web using the Language Markup and Experimental Design Software (LMEDS; Mahrt 2016). The task of the first group of participants was to listen to spontaneously spoken utterances and to indicate in the transcriptions displayed on a computer screen where some sort of juncture could be heard. More specifically, participants were asked to click on the words that in their opinion precede some sort of juncture in the audio recording. No additional explanations about the type of juncture they should look for were provided. The purpose of the methodology is that listeners find their own internal criteria for chunking the stream of speech into units larger than words. The participants could listen to the recordings of corresponding utterances two times. Listening to and judging the list of 99 utterances took about 40–60 min. The participants were encouraged to take short breaks. Based on the concluding questionnaire, the participants could only poorly guess the underlying research question. To emphasize on the highly spontaneous nature of our speech materials, we note that a number of participants complained about the bad manners of spontaneously spoken Estonian.

The task of the second group of participants was to read transcriptions of spontaneously spoken utterances. As the utterances come from the phonetic corpus of spontaneous speech, it was not possible to provide punctuation marks. The participants were asked to indicate where the words form some sort of grouping that is larger than a word but smaller than the whole utterance. Similar to the task of listening, they were asked to click on the word that is at the end of some sort of a word grouping. As with the listening task, no further examples or instructions for defining the groups of words were provided. On average, the reading and rating of 99 sentences took about 28 min. Based on the concluding questionnaire, few participants were convinced that the task was to put the boundary marks at places of punctuation as they occur in written language.

3.1.4 Analysis

Participants rated consecutive pairs of words for the presence of word boundary by clicking on the words. When a word was clicked on, then it was scored with 1. Conversely, when it was not clicked on, it received a score of 0. As such, the response variable indicates the number of occurrences of boundary marks and can

be modelled with the parameters of a probability distribution. For this, we devised again a GLMER analysis that estimated the probability of the boundary mark as a function of prosodic variables (Section 2.1.2) and linguistic variables (Section 2.1.3). Thus, the relevant parts of the model included the semantico-syntactic variables (i.e. clause boundary, phrase boundary, bigram frequency, and the trigram frequency) that were the dependent variables of the corpus study.

The aim of the analysis was to test the contribution of *Syllable Duration*, *Intensity Difference*, *Declination Estimate*, and *Relative F0* on the perception of prosodic breaks as modulated by the distributions of *Clause Boundaries* and *Phrase Boundaries*. Due to too few occurrences, it was not possible to include *Pause Duration* (0.07%, $N = 357$) in the models of boundary perception. Therefore, the GLMER analyses included interactions between the continuous prosodic variables (i.e. *Syllable Duration*, *Intensity Difference*, *Declination Estimate*, *Relative F0*) and categorical variables (i.e. *Clause Boundaries* and *Phrase Boundaries*); (8 interaction terms altogether). The frequencies of bigrams and trigrams were tested as the main effects. The preliminary observation of model fits indicated that adding the interactions between the prosodic variables and the collocation frequencies did not enhance the model fit (based on the Akaike information criterion (AIC) and distribution of residuals). Similarly, the preliminary diagnostics of the model fits revealed that the two separate regression analyses by processing mode (listening vs reading) yielded a better fit than a single model including the mode as one of the explanatory variables. Thus, the effects of the explanatory variables on the perception of prosodic breaks in the two modes were estimated in two separate GLMER analyses. For the regression analyses, the duration of syllables and the frequency of bigrams and trigrams were logarithmically transformed with the base of 10. All continuous variables (i.e. *Bigram Frequencies*, *Trigram Frequencies*, *Syllable Duration*, *Intensity Difference*, *Declination Estimate*, and *Relative F0*) were z-scored before the submission to the regression analyses. The scaling and centring procedures enable us to compare the estimates within and across the two models.

In addition, the GLMER fits were specified for the exposure variable that was the number of listeners per excerpt. The random effects structure included random slopes for listeners by all of the dependent variables because we reasoned that listeners are highly likely to vary in their sensitivity to the syntactic structure and the distribution of prosodic information. In addition, the model contained the random slopes for speakers by the prosodic variables because the 396 utterances were spoken by 10 different speakers. These speakers constitute a categorical variable which is likely to contain the speaker-specific systematically varying prosodic information. For obtaining a better model fit, the correlation terms between the intercepts and slopes were excluded. As in the corpus study, the mixed effects regression analyses were conducted with the help of the *lme4* package (Bates et al. 2015) in R (Core Team 2019). The converging model fit was obtained by using the *optimx* optimizer (Nash 2014, Nash and Varadhan 2011). The significance tests were computed with the *jtools* package (Long 2017). The GLMER model diagnostics was conducted with the package *DHARMa* (Dunn and Smyth 1996, Gelman and Hill 2006).

3.2 Results

The two categorical (*Clause Boundary*, *Phrase Boundary*) and six continuous variables (*Bigram Frequencies*, *Trigram Frequencies Syllable Duration*, *Intensity Difference*, *Declination Estimate*, and *Relative F0*) were predicted to contribute to the perception of prosodic breaks. Before running the regression analyses, we estimated the correlations between the explanatory variables by calculating Pearson's r values (Table 4). The categorical variables *Clause Boundary* and *Phrase Boundary* were treated as numeric variables in the analysis of correlations.

Table 4 indicates that the relationships between the explanatory variables do not reach strong degree of correlation (0.7 or higher). Thus, the multiple regression analysis is suited to examine the simultaneous effects of prosodic and non-prosodic variables on the perception of prosodic breaks.

The GLMER analysis estimated the probability of the perception of a boundary, given the variation of prosodic and non-prosodic variables. We expected the greater contribution of prosodic variables in the

Table 4: Pearson's *r* correlations between the nine explanatory variables of boundary perception

	Clause B.	Phrase B.	Bigram Freq.	Trigram Freq.	Decl. Est.	Syl. Dur.	Pause Dur.	Int. Dif.
Phrase B.	0.52							
Bigram Freq.	0.23	0.01						
Trigram Freq.	−0.12	−0.12	0.61					
Decl. Est.	−0.13	−0.29	0.09	−0.03				
Syl. Dur.	−0.46	−0.20	0.21	0.32	−0.03			
Pause Dur.	0.50	0.52	0.06	−0.22	−0.29	−0.19		
Int. Dif.	−0.15	−0.09	0.02	−0.01	0.14			
Rel. F0	−0.13	0.01	−0.12	0.10	−0.25	0.27	−0.02	−0.29

listening mode than in reading mode, that is larger values for the model estimates. In particular, the perception of prosodic breaks was expected at the clause and phrase boundaries and at the boundaries of collocations of greater frequency. Moreover, we expected significant interactions between the syntactic variables (i.e. *Clause Boundary*, *Phrase Boundary*) and the prosodic variables (i.e. *Syllable Duration*, *Intensity Difference*, *Declination Estimate*, and *Relative F0*) because we have suggested that the prosodic discontinuities are better audible at the boundaries of structurally defined units (e.g. clause boundaries, phrase boundaries, collocations). Very specifically, to see support for our expectations, the estimates of constituent boundaries within the interactions with prosodic variables should be significantly greater than zero and greater for the listening mode than the reading mode (i.e. to the further right of the zero line in Figure 2).

Expectedly, the GLMER analysis of boundary perception indicates that prosodic discontinuities have a larger effect on the perception of breaks in the listening mode than in the reading mode (for the significant effects across the two models, see Figure 2). The analysis indicated for listening that the perception of prosodic breaks was significantly influenced by the main effects of *Bigram Frequency* (Est. = −0.36, CI = (−0.49, −0.23), $z = -5.39$, $p < 0.001$), *Syllable Duration* (Est. = 0.9, CI = (0.68, 1.12), $z = 8.02$, $p < 0.001$), *Clause Boundary* (Est. = 3.05, CI = (2.7, 3.39), $z = 17.45$, $p < 0.001$) and *Phrase Boundary* (Est. = −0.42, CI = (−0.65, −0.18), $z = -3.43$, $p < 0.001$). Importantly, *Syllable Duration* (Est. = −0.7, CI = (−0.92, −0.48), $z = -6.13$, $p < 0.001$), *Intensity Difference* (Est. = −0.87, CI = (−1.16, −0.57), $z = -5.77$, $p < 0.001$), and *Declination Estimate* (Est. = 0.35, CI = (0.14, 0.56), $z = 3.25$, $p < 0.001$) contributed to the perception of chunk boundaries within the interactions with clause boundaries. Similarly, *Syllable Duration* (Est. = 0.25, CI = (0.05, 0.45), $z = 2.49$, $p < 0.05$), *Intensity Difference* (Est. = 0.71, CI = (0.43, 1.0), $z = 4.93$, $p < 0.001$), and *Relative F0* (Est. = −0.30, CI = (−0.50, −0.10), $z = -2.90$, $p < 0.001$) influenced the boundary perception within the interactions with phrase boundaries. There were no significant main effects of *Trigram Frequency*, *Intensity Difference*, *Declination Estimate*, and *Relative F0* and no significant interactions between *Clause Boundary* and *Relative F0* and between *Phrase Boundary* and *Declination Estimate*.

For the reading, the analysis reveals that the perception of a break is influenced by the main effects of *Bigram Frequency* (Est. = −0.37, CI = (−0.51, −0.22), $z = -5.01$, $p < 0.001$), *Syllable Duration* (Est. = 0.38, CI = (0.2, 0.56), $z = 4.16$, $p < 0.001$), *Clause Boundary* (Est. = 3.98, CI = (3.57, 4.39), $z = 19.00$, $p < 0.001$) and *Phrase Boundary* (Est. = −0.58, CI = (−0.82, −0.33), $z = -4.57$, $p < 0.001$). Interestingly, also some prosodic variables played a role in the marking of boundaries. For instance, *Syllable Duration* (Est. = −0.59, CI = (−0.84, −0.34), $z = -4.67$, $p < 0.001$) and *Declination Estimate* (Est. = 0.35, CI = (0.11, 0.59), $z = 2.86$, $p < 0.001$) affected the boundary perception within the interactions with clause boundaries. In addition, *Declination Estimate* (Est. = −0.40, CI = (−0.65, −0.15), $z = -3.14$, $p < 0.001$) contributed to the boundary perception in the reading mode within the interaction with phrase boundaries. The main effects of *Trigram Frequency*, *Intensity Difference*, *Declination Estimate*, and *Relative F0* and the interactions between *Clause Boundary* and *Intensity Difference*, between *Clause Boundary* and *Relative F0*, between *Clause Boundary* and *Syllable Duration*, between *Phrase Boundary* and *Intensity Difference*, and between *Phrase Boundary* and *Relative F0* were not significant.

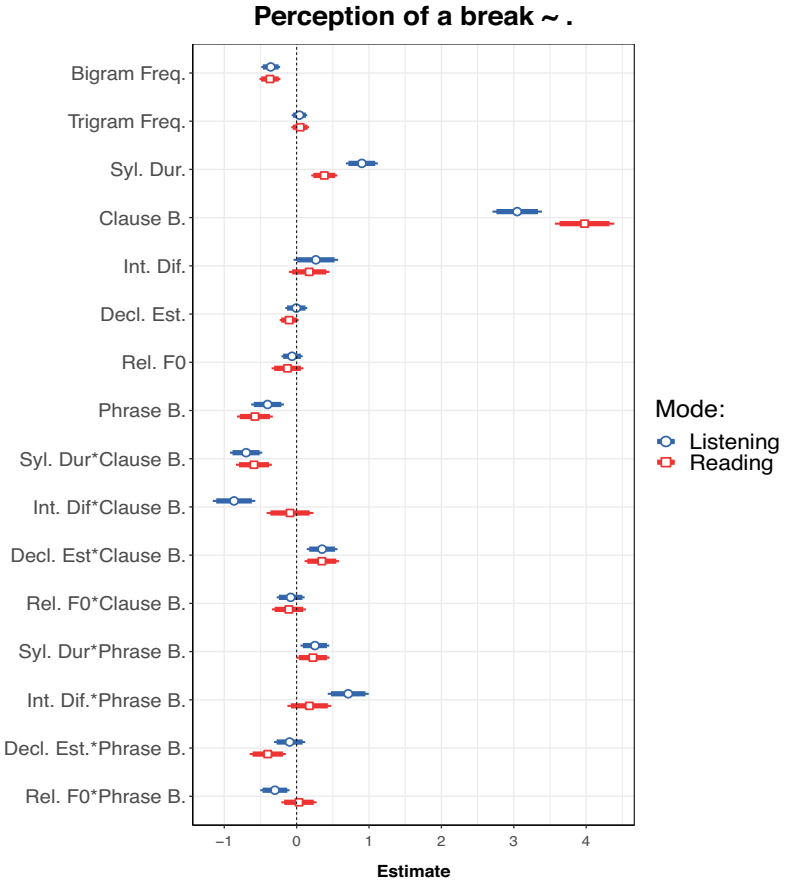

Figure 2: Contributions of collocation frequencies (*Bigram Frequency*, *Trigram Frequency*) and prosodic discontinuities (*Pause Duration*, domain-specific *Declination Estimate*, word-final *Syllable Duration*, *Intensity Difference*, *Relative F0*) to the perception of prosodic breaks. The effect sizes were obtained separately for the listening mode (blue circles and blue lines) and the reading mode (red squares and red lines). The points illustrate the model estimates and the lines the 95% confidence intervals around the estimates.

3.3 Interim discussion

The perceptual study of language chunking based on the auditive and written presentation of spontaneous utterances has yielded expected patterns of results. First, the strongest influence on the perception of chunk boundaries was exerted by the clause boundaries (Figure 2). The analysis confirmed that a boundary was perceived more likely in the presence of clause boundary than in the absence of a clause boundary. The effect of clause boundaries was stronger in the reading mode than in the listening mode (see the red square and the red line of the term "Clause B." in Figure 2). The next strongest effect on the perception of boundaries came from the duration of word-final syllables. The longer was the syllable duration, the likelier was the perception of a boundary. As expected, the effect of syllable duration was stronger in the listening mode than in the reading mode (see the blue circle and the blue line of the term "Syl. Dur." in Figure 2).

Second, the effects of clause and phrase boundaries interacted significantly with the prosodic variables. While most of the interaction terms involving prosodic variables reached significance in the listening mode, only *Syllable Duration* and *Declination Estimate* turned out significant in the reading mode. In particular, the significant interactions indicated for listening that the increasing syllable duration had a particularly strong effect on boundary perception in the absence of clause boundaries but the effect was somewhat smaller in the presence of clause boundaries (the estimates in Figure 2 are supplemented with the model predictions in Figure 3a). The increasing intensity difference contributed to the perception of boundaries in the absence of clause boundaries. However, the effect of intensity was opposite in the presence of clause

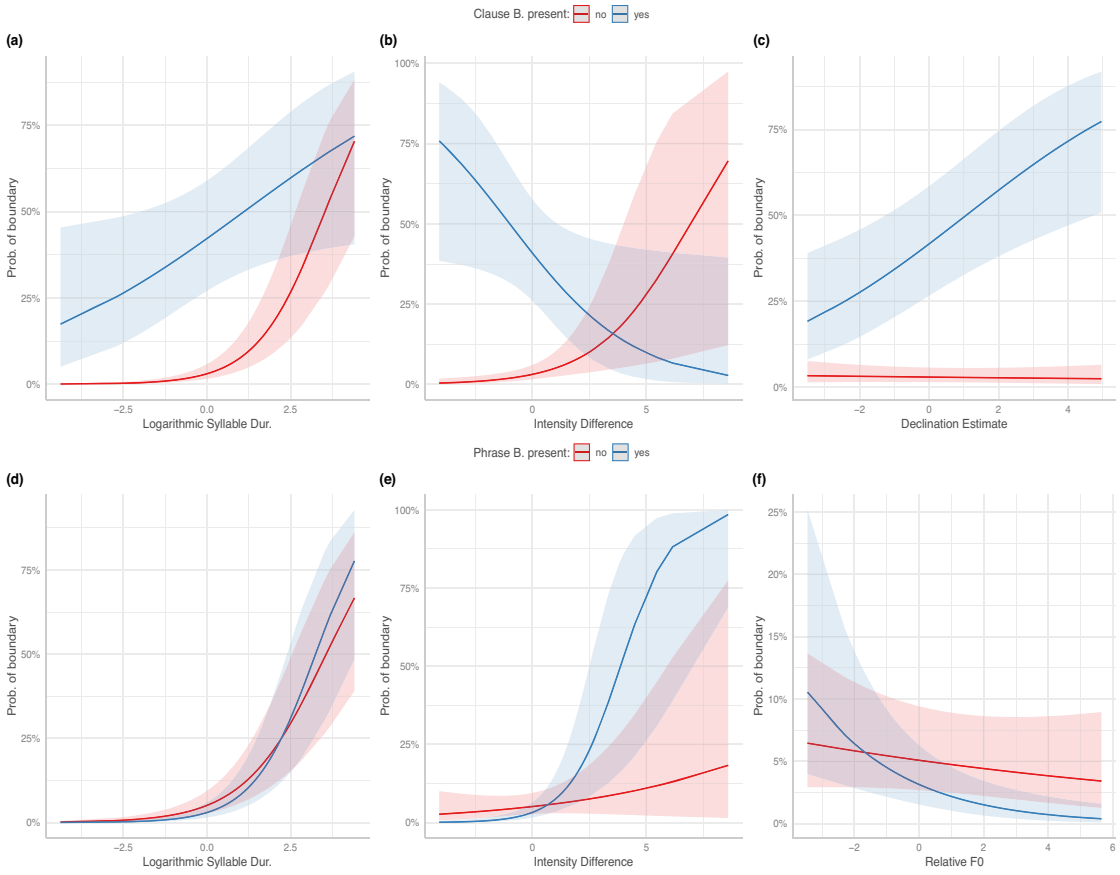

Figure 3: The predictions of significant interactions supplementary to Figure 2. The plots indicate the probability of boundary perception as a function of increasing syllable duration (a), intensity difference (b), level of F0 declination (c) for clause boundaries and the probability of boundary perception as a function of increasing syllable duration (d), intensity difference (e), and relative F0 (f) for phrase boundaries. The shadowed areas around the lines represent 95%-confidence intervals of the estimates.

boundaries (see Figure 3b for the model predictions). The influence of estimated F0 affected the perception of boundaries only in the presence of clause boundaries such that the higher was the estimated F0, the likelier was the boundary perception (Figure 3c). Conclusively, the results suggest that while syllable duration and intensity difference are in the trading relationship with clause boundaries, the effect of declination estimate depends on the presence of the clause boundaries. In other words, either increasing syllable duration and increasing intensity difference or a clause boundary can independently trigger the perception of chunk boundary. Differently, though, the level of F0 declination has an effect only when the clause boundary is present.

Also, phrase boundaries interacted with prosodic information in the perception of chunk boundaries. For example, the probability of boundary perception increased together with increasing syllable duration and the increase in the probability of boundary perception was somewhat sharper when phrase boundaries were present than when they were absent (Figure 3d). Furthermore, the increase of boundary perception as a function of increasing intensity difference was somewhat greater in the presence of phrase boundaries, than in the absence of phrase boundaries (Figure 3e). Finally, the phrase boundaries modulated the perception of relative F0 such that the probability of boundary perception increased together with the increasing word-final syllable in the presence of phrase boundaries but not in the absence of phrase boundaries (Figure 3f). In other words, the probability of boundary perception increased together with the increasing syllable duration with a very little influence from the presence of the phrase boundaries, whereas the effects of intensity and F0 occurred only in the presence of phrase boundaries. Thus, there was an independent effect of syllable duration but effects of intensity and F0 were modulated by the presence of phrase boundaries.

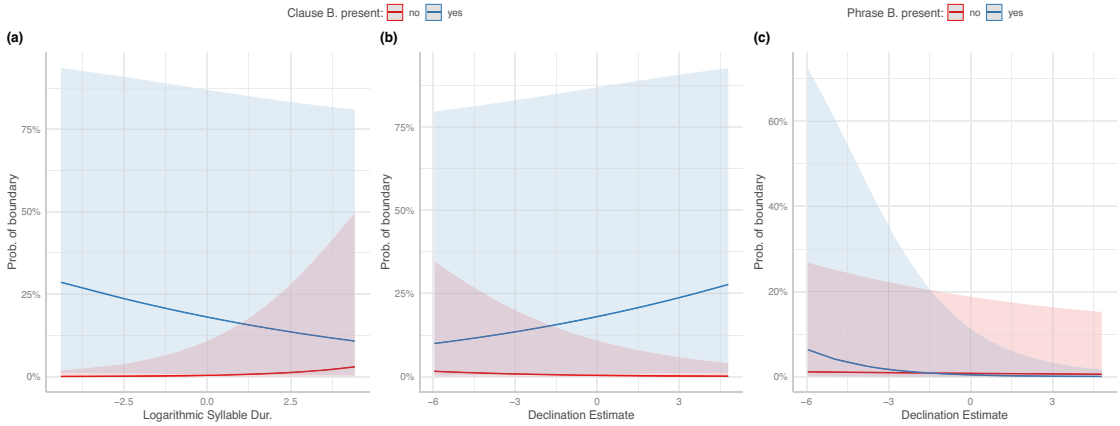

Figure 4: The predictions of significant interactions supplementary to Figure 2. The plots indicate the probability of boundary perception as a function of increasing syllable duration, and level of F0 declination for clause boundaries (a and b, respectively) and the probability of boundary perception as a function of increasing level of F0 declination for phrase boundaries (c). The shadowed areas around the lines represent 95%-confidence intervals of the estimates.

In the reading mode, the chunk boundaries were mainly detected at the clause boundaries, whereas the phrase boundaries had a negative effect on boundary marking. Interestingly, some of the interactions with prosodic information still turned out significant. In particular, the analysis indicated that the probability of the boundary perception increased together with the increasing syllable duration only in the absence of clause boundary (Figure 4a). Similarly, the estimated F0 influenced the perception of chunk boundary only when the clause boundary was present. In conjunction with clause boundaries, the probability of boundary perception decreased as the estimated F0 increased (Figure 4b). Finally, the perception of phrase boundaries as chunk boundaries was somewhat affected by the level of estimated declination. Namely, the probability of boundary perception increased together with decreasing estimated F0 only when phrase boundaries were present but not when the phrase boundaries were absent (Figure 4c). However, as both Figures 2 and 4 indicate, the effects are rather small and the confidence intervals rather wide.

Finally, as a novelty, the effect of bigrams occurred in both listening and reading modes. In particular, the probability of boundary perception decreased as the frequency of a bigram increased. Notably, the increasing bigram frequency was taken to indicate a greater likelihood of a collocation boundary. For the effect of collocation boundaries on the boundary perception, we expected that the boundary perception increases together with the increasing collocation frequency. However, the results show that the perception of a chunk boundary is more likely at the boundaries of rather infrequent than frequent collocations.

4 General discussion

The aim of the study was to investigate whether the perception of prosodic discontinuities as prosodic breaks is modulated by the structurally defined positions such as boundaries of constituents and collocations. By large, our results are consistent with an idea that the structurally defined boundaries make the prosodic discontinuities better audible.

In the study of spontaneous speech production, we predicted that longer duration of pauses, longer duration of word-final syllables, that is *pre-boundary lengthening*, greater intensity difference between the word-initial and word-final syllables, that is *intensity drop*, higher starting level of F0 declination, that is *pitch reset*, and higher relative F0 maxima, that is *high boundary tone* correspond with the boundaries of syntactic clauses and phrases, and with the boundaries of bigrams and trigrams. We did not find support for these expectations. The results of the corpus study suggest a rather weak correspondence between the prosodic discontinuities and the syntactic information in naturally spoken spontaneous utterances. The

general linear mixed effects regression analyses even detected an effect in an opposite direction where the shorter than longer syllable durations indicated a higher probability of a clause boundary. We argue that this effect is driven by the lengthened syllables in the utterance-final position that were not accompanied by a clause boundary. The utterance-final lengthening that occurs within the syntactic clauses indicates pragmatic and/or production constraints that may be far more prevalent for the spontaneous interactions than the need to mark the clause boundaries. Altogether, the findings suggest that the syntactic organization of sentences might exert only weak control over spontaneous production of prosody. Thus, a lot of prosodic variation in spontaneous conversations remains unaccounted by the structural features of utterances.

In the perception study, we investigated to what degree the distribution of prosodic information might map onto the syntactically and probabilistically defined elements of language (the syntactic constituents and frequent two-word and three-word sequences, respectively). For the perception of prosodic breaks, we predicted that listeners are more sensitive to longer duration of pauses and syllables, greater intensity difference within words, higher F0 declination, and higher F0 peaks at the edges of syntactic constituents and frequent collocations whilst listening and reading the spontaneous utterances than whilst only reading them. In other words, we expected that the effects of pre-boundary lengthening, intensity drop, pitch reset, and high boundary tone on boundary perception are augmented by the structurally meaningful points of utterances when listeners have access to acoustic information. In contrast, the effects of prosodic information were expected to diminish or disappear when the acoustic information is inaccessible. By large, the results are consistent with these expectations.

In particular, the perception of pre-boundary lengthening and intensity drop as prosodic breaks was less affected by the presence of constituent boundaries than the perception of melodic discontinuities (e.g. pitch reset and high boundary tone). The perception of pre-boundary lengthening and intensity difference as a prosodic break was less affected by the presence of constituent boundaries. This result reflects well the earlier phonetic findings showing that the contribution of duration and intensity in the perception of breaks is stronger than the contribution of pitch cues (see e.g. Hawthorne 2018, Peters 2005, Petrone et al. 2017, Streeter 1978, Swerts et al. 1994). Importantly, the results of our study provide that pitch reset and high boundary tones require a boost from the syntactic boundaries to be perceived as prosodic phrase boundaries. In addition, the importance of prosodic information decreased when participants did not have access to the acoustic information (i.e. reading mode). While the effect of clause boundaries was even greater in reading mode than in listening mode, the participants of a reading experiment still uncovered elements of language that correspond quite well with the notion of prosodic coherence. In particular, the perceptual units they identified showed features of pre-boundary lengthening and pitch reset. This outcome might indicate that readers of the transcriptions of spontaneous speech marked boundaries not only at the clause boundaries but probably at the boundaries of other types of elements as well (e.g. disclosures and interjections). If so, then the results confirm that clauses together with some other types of language units systematically trigger prosodic discontinuities which are detectable based on text only. This, in turn, corroborates the idea in Cole et al. (2010) that the clause boundaries might be perceived as prosodic breaks because they are frequently accompanied by the lengthening of final syllables. The fact that pitch reset and pre-boundary lengthening turned out to be important also for the chunking of written language underscores the importance of prosodic information in production and perception of speech chunks.

In line with the speech processing account in Dahan and Ferreira (2019), we have proposed that the function of rhythmic and melodic discontinuities in real-time processing of speech is to activate the set of syntactic and semantic structures that are possibly generating a particular speech signal. The activation of these structures then guides the mental attention away from the rhythmic and melodic discontinuities that do not converge with the characteristics of predicted structures. The empirical implication of this proposal is that any sort of structurally defined language chunk, either syntactically defined ones such as syntactic constituents or frequency-based ones such as bigrams should augment the perception of prosodic discontinuities as prosodic phrase boundaries. With this regard, our results confirm the influence from the clause boundaries but not the influence from collocation boundaries. Admittedly, the word pairs and word triplets detected in our materials were rather infrequent in the corpus of written language (Raudvere and Uiboaed 2018). The use of the corpus of spontaneous Estonian (Lippus et al. 2016) would not helped either because it

is not large enough. Often, the substantives in our selected excerpts constituted a single occurrence in that corpus. The results on collocation boundaries should be thus taken with caution. Moreover, the future studies should incorporate methodologies that go beyond a meta-linguistic judgement tasks (e.g. the chunking task).

The study was limited to the investigation of syntactic constituent structure and the probabilistic organization of collocations. As was discussed in Section 1, the perception of prosodic discontinuities may further be mediated by the prosodic-metric structure as well. Another type of structural representation of prosody that might be useful for processing spoken language is the cyclical models of intonational tunes (e.g. Pierrehumbert 1980). These models provide predictions for the upcoming breaks because they establish the allowed sequences of different intonational pitch accents, phrase accents, and edge tones for a particular language. For example, when a certain type of a pitch accent is only allowed together with a type of edge tone, then upon hearing this pitch accent, the occurrence of a break based on the tone concurrence rules is already highly predictable. The cyclical model of Estonian intonation, however, only weakly constrains the combinations between the different pitch accents and edge tones (Asu 2005, Asu and Nolan 2007). The most frequent pitch accent in Estonian is the falling pitch accent that usually combines with the low edge tone (Asu and Nolan 1999). Thus, the Estonian intonation grammar has a rather weak predictive power for the co-occurrences of intonational tunes. A language with much richer tonal inventory and a greater number of rules could inform us whether the abstract representation of prosody would modulate the perception of prosodic breaks to a similar degree of syntactic constituent structure.

5 Conclusion

The corpus study demonstrated for Estonian that clause and phrase boundaries in spontaneously spoken utterances correspond only weakly with prosodic discontinuities (i.e. pre-boundary lengthening, intensity drop, pitch reset, and high boundary tones). The perception study showed that the prosodic discontinuities differ in the degree of which they are perceived as prosodic breaks. While intensity drop and pre-boundary lengthening established a trading relationship with constituent boundaries, the perception of pitch reset and high boundary tone as prosodic breaks depended on the presence of syntactic boundaries. These findings are consistent with the idea that listeners might not notice or they might rapidly forget a great deal of prosodic information in spontaneous interactions when these acoustic details are not accompanied by some sort of structurally meaningful information. As such, the study demonstrates the importance of structural as well as prosodic information in speech chunking. Moreover, the proposal is that the prosodic information first activates the structural information but this needs to be captured in the future studies.

Acknowledgments: The authors are extremely grateful to the volunteers who contributed to our first experiment and to participants of the second experiment. The authors appreciate the support and motivation that they have received from the audiences of Societas Linguistica Europaea meetings (i.e. the SLE 2018 where their presentation got the third prize and the SLE 2019).

Funding information: This work was financed by the Estonian Research Council grant (PSG671) and by the European Union through the European Regional Development Fund (Centre of Excellence in Estonian Studies) and by the Fritz Thyssen Stiftung in Germany (10.18.2.040SL, "Planning sentences and sentence intonation cross-linguistically").

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission. Conceptualization of the experiment – NO and PT; administration of the experiments – NO; validation of the results – NO; visualization of the results – NO; writing, original draft – NO; writing, review, and editing – NO and PT.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The datasets generated and analysed during the current study are available in the OSF repository, https://osf.io/57k3c/. The datasets (speech recordings) analysed during the current study are available from the corresponding author on reasonable request.

References

Asu, Eva Liina. 2005. "Towards a phonological model of Estonian intonation." In: *Proceedings of the Second Baltic Conference on Human Language Technologies, Tallinn 4–5 May 2005*, edited by Langemets, Margit and Priit Penjam. p. 95–100. Tallinn: Tallinn University of Technology and Institute of the Estonian Language.

Asu, Eva Liina and Francis Nolan. 1999. "The effect of intonation on pitch cues to the Estonian quantity contrast." In: *Proceedings of the 14th International Congress of Phonetic Sciences*, edited by Ohala, J., p. 1873–6. San Francisco: University of California.

Asu, Eva Liina and Francis Nolan. 2007. "The analysis of low accentuation in Estonian." *Language and Speech* 50(4), 567–88.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting linear mixed-effects models using lme4." *Journal of Statistical Software* 67(1), 1–48.

Beckman, Mary E. 1986. *Stress and non-stress accent. Vol. 7 of Netherlands phonetics archives*. Holland, Dordrecht: Foris Publication.

Beckman, Mary E. and Jan Edwards. 1990. "Lengthenings and shortenings and the nature of prosodic constituency." In: *Papers in Laboratory Phonology: Volume 1, Between the Grammar and Physics of Speech*, edited by John Kingston and Mary E Beckman, p. 152–78.

Beňuš, Š., U. D. Reichel, and K. Mády. 2014. "Modelling accentual phrase intonation in Slovak and Hungarian." In: *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*, edited by Veselovská, L. and M. Janebová, p. 677–89. Olomouc: Palacký University.

Berkovits, Rochele. 1994. "Durational effects in final lengthening, gapping, and contrastive stress." *Language and Speech* 37(3), 237–50. pMID: 7861912. doi: https://doi.org/10.1177/002383099403700302.

Bögels, S., H. Schriefers, W. Vonk, and D. J. Chwilla. 2011. "The role of prosodic breaks and pitch accents in grouping words during on-line sentence processing." *Journal of Cognitive Neuroscience* 23(9), 2447–67.

Bögels, S., H. Schriefers, W. Vonk, D. J. Chwilla, and R. Kerkhofs. 2010. "The interplay between prosody and syntax in sentence processing: the case of subject- and object-control verbs." *Journal of Cognitive Neuroscience* 22(5), 1036–53.

Bögels, S., H. Schriefers, W. Vonk, D. J. Chwilla, and R. Kerkhofs. 2013. "Processing consequences of superfluous and missing prosodic breaks in auditory sentence comprehension." *Neuropsychologia* 51(13), 2715–28. https://www.sciencedirect.com/science/article/pii/S0028393213002923.

Blaauw, Eleonora. 1994. "The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech." *Speech Communication* 14(4), 359–75. https://www.sciencedirect.com/science/article/pii/0167639394900280.

Bock, Kathryn, David E. Irwin, and Douglas J. Davidson. 2004. "Putting first things first." In: *The interface of language, vision, and action: Eye movements and the visual world*, edited by Henderson, J. and F. Ferreira, p. 249–78. New York, NY, US: Psychology Press.

Bock, K. and Willem Levelt. 1994. "Language production. grammatical encoding." In: *Handbook of psycholinguistics*, edited by Gernsbacher, M. A., p. 945–84. San Diego: Academic Press, Chapter 29.

Boersma, P. and D. Weenink. 2020. Praat: Doing phonetics by computer [computer program]. Version 6.1.09, retrieved 26 January 2020 from http://www.praat.org/.

Borrelli, Dario, Gabriela Gongora Svartzman, and Carlo Lipizzi. June 2020. "Unsupervised acquisition of idiomatic units of symbolic natural language: An n-gram frequency-based approach for the chunking of news articles and tweets." *PLOS ONE* 15, 1–18. doi: https://doi.org/10.1371/journal.pone.0234214.

Bürki, Audrey. 2018. "Variation in the speech signal as a window into the cognitive architecture of language production." *Psychonomic Bulletin and Review* 25(6), 1973–2004. doi: https://doi.org/10.3758/s13423-017-1423-4.

Buxó-Lugo, Andrés and Duane G. Watson. 2016. "Evidence for the influence of syntax on prosodic parsing." *Journal of Memory and Language* 90, 1–13, http://www.sciencedirect.com/science/article/pii/S0749596X16000231.

Bybee, Joan L. 2002. "Sequentiality as the basis of constituent structure." *Vol. 53 of typological studies in language*. 109–34. Amsterdam: John Benjamins Publishing Company.

Cambier-Langeveld, Tina. 1997. "The domain of final lengthening in the production of Dutch." *Linguistics in the Netherlands* 14, 13–24.

Carpenter, Patricia A. and Marcel A. Just. 1989. *The role of working memory in language comprehension.* p. 31–68. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Cho, Taehong and Patricia A. Keating. 2001. “Articulatory and acoustic studies on domain-initial strengthening in Korean.” *Journal of Phonetics* 29(2), 155–90. https://www.sciencedirect.com/science/article/pii/S0095447001901317.
Christiansen, Morten H. and Nick Chater. 2016. “The now-or-never bottleneck: A fundamental constraint on language.” *Behavioral and Brain Sciences* 39, e62.
Christophe, Anne, Sharon Peperkamp, Christophe Pallier, Eliza Block, and Jacques Mehler. 2004. “Phonological phrase boundaries constrain lexical access I. Adult data.” *Journal of Memory and Language* 51(4), 523–47. https://www.sciencedirect.com/science/article/pii/S0749596X04000816.
Clark, Andy. 2003. “Whatever next? predictive brains, situated agents, and the future of cognitive science.” *The Behavioral and Brain Sciences* 36(3), 181–204.
Clifton, Charles Jr, Katy Carlson, and Lyn Frazier. Oct. 2006. “Tracking the what and why of speakers’ choices: prosodic boundaries and the length of constituents.” *Psychonomic Bulletin & Review* 13(5), 854–61.
Cole, Jennifer, Yoonsook Mo, and Soondo Baek. 2010. “The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech.” *Language and Cognitive Processes* 25(7–9), 1141–77.
Cooper, W. E. and J. Paccia-Cooper. 1980. *Syntax and speech*. Cambridge, MA: Harvard University Press.
Cooper, W. E. and J. M. Sorensen. 1981. *Fundamental frequency in sentence production*. New York: Springer-Verlag.
Cutler, Anne. 1976. “Phoneme-monitoring reaction time as a function of preceding intonation contour.” *Perception and Psychophysics* 20(1), 55–60.
Cutler, Anne, Delphine Dahan, and Wilma van Donselaar. 1997. “Prosody in the comprehension of spoken language: A literature review.” *Language and Speech* 40(2), 141–201.
Dahan, Delphine. 2015. “Prosody and language comprehension.” *WIREs Cognitive Science* 6(5), 441–52. https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1355.
Dahan, Delphine and Fernanda Ferreira. 2019. “Language comprehension: Insights from research on spoken language.” In: *Human language: from genes and brains to behavior*, edited by Hagoort, P., p. 21–33. Cambridge, MA: MIT Press.
Denham, S. L. and I. Winkler. 2006. “The role of predictive models in the formation of auditory streams.” *Journal of Physiology* 100, 154–70.
Duez, Danielle. 1985. “Perception of silent pauses in continuous speech.” *Language and Speech* 28(4), 377–88.
Dunn, Peter K. and Gordon K. Smyth. 1996. “Randomized quantile residuals.” *Journal of Computational and Graphical Statistics* 5(3), 236–44. http://www.jstor.org/stable/1390802.
Erelt, Mati and Helle Metslang (Eds.), 2017. *Eesti keele süntaks* [*eng. Estonian Syntax*]. *No. 3 in Eesti keele varamu*. Tartu: Tartu Ãœlikooli Kirjastus.
Ferreira, Fernanda and Hossein Karimi. 2015. *Prosody, performance, and cognitive skill: evidence from individual differences*. p. 119–32. Cham: Springer International Publishing. doi: https://doi.org/10.1007/978-3-319-12961-7_7.
Féry, Caroline and Shinichiro Ishihara. 2009. “How focus and givenness shape prosody.” In: *Information structure: theoretical, typological, and experimental perspectives*, edited by Zimmermann, Malte and Caroline Féry, p. 36–63. Oxford: Oxford University Press.
Fon, Janice, Keith Johnson, and Sally Chen. Mar 2011. “Durational patterning at syntactic and discourse boundaries in mandarin spontaneous speech.” *Language and speech* 54, 5–32.
Andrew Fromkin. 1971. “The non-anomalous nature of anomalous utterances.” *Language* 47, 27.
Gelman, Andrew and Jennifer Hill. 2006. “Data analysis using regression and multilevel/hierarchical models.” *Analytical methods for social research*. Cambridge, UK: Cambridge University Press.
Grosjean, François, Lysiane Grosjean, and Harlan Lane. 1979. “The patterns of silence: Performance structures in sentence production.” *Cognitive Psychology* 11(1), 58–81. http://www.sciencedirect.com/science/article/pii/0010028579900045.
Hawthorne, Kara. 2018. “Prosody-driven syntax learning is robust to impoverished pitch and spectral cues.” *The Journal of the Acoustical Society of America* 143(5), 2756–67. doi: https://doi.org/10.1121/1.5031130.
Himmelmann, Nikolaus P., Meytal Sandler, Jan Strunk, and Volker Unterladstetter. 2018. “On the universality of intonational phrases: a cross-linguistic interrater study.” *Phonology* 35(2), 207–45.
Keating, P., T. Cho, C. Fougeron, and C.-S. Hsu. 2003. “Domain-initial strengthening in four languages.” In: *Laboratory phonology VI: Phonetic interpretation*, edited by Local, J., R. Ogden, R. Temple, p. 145–63. Cambridge: Cambridge University Press.
Keating, P. and S. Shattuck-Hufnagel. August 2002. “A prosodic view of word form encoding for speech production.” *UCLA Working Papers in Phonetics* 101, 112–56.
Keating, Patricia A. 2006. “Phonetic encoding of prosodic structure.” *Speech production: Models, phonetic processes and techniques*, edited by J. Harrnington and M. Tabain, p. 167–86. New York: Psychology Press.
Kentner, Gerrit and Caroline Féry. 2013. “A new approach to prosodic grouping.” *The Linguistic Review* 30(2), 277–311. https://www.degruyter.com/view/journals/tlir/30/2/article-p277.xml.
Kerkhofs, Roel, Wietske Vonk, Herbert Schriefers, and Dorothee J. Chwilla. Aug 2008. “Sentence processing in the visual and auditory modality: do comma and prosodic break have parallel functions?.” *Brain Research* 1224, 102–18.
Klatt, Dennis H. 1975. “Vowel lengthening is syntactically determined in a connected discourse.” *Journal of Phonetics* 3(3), 129–40. https://www.sciencedirect.com/science/article/pii/S0095447019313609.
Konopka, Agnieszka E. and Antje S. Meyer. 2014. “Priming sentence planning.” *Cognitive Psychology* 73, 1–40.

Kraljic, Tanya and Susan E. Brennan. 2005. "Prosodic disambiguation of syntactic structure: For the speaker or for the addressee?." *Cognitive Psychology* 50(2), 194–231. https://www.sciencedirect.com/science/article/pii/S0010028504000702.

Krivokapić, Jelena. 2007. "Prosodic planning: Effects of phrasal length and complexity on pause duration." *Journal of Phonetics* 35(2), 162–79. http://www.sciencedirect.com/science/article/pii/S0095447006000180.

Ladd, D. Robert. 1988. "Declination 'reset' and the hierarchical organization of utterances." *The Journal of the Acoustical Society of America* 84, 530–44.

Ladd, D. Robert. 2008. "Intonational phonology." *Cambridge studies in linguistics*. Vol. 119. Cambridge: Cambridge University Press.

Langus, Alan, Erika Marchetto, Ricardo Augusto Hoffmann Bion, and Marina Nespor. 2012. "Can prosody be used to discover hierarchical structure in continuous speech?." *Journal of Memory and Language* 66(1), 285–306. https://www.sciencedirect.com/science/article/pii/S0749596X11001021.

Lee, Eun-Kyung, Sarah Brown-Schmidt, and Duane G. Watson. Dec. 2013. "Ways of looking ahead: hierarchical planning in language production." *Cognition* 129(24045002), 544–62. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3909534/.

Lehiste, Ilse. 1972. "The timing of utterances and linguistic boundaries." *Journal of the Acoustical Society of America* 51, 2018–24.

Lehiste, Ilse. 1973. "Phonetic disambigation of syntactic ambiguity." *Glossa* 7, 107–22.

Levelt, W. J. M. 1989. *Speaking: From intention to articulation*, Cambridge, MA: MIT Press.

Lindström, L. 2006. "Infostruktuuri osast eesti keele sõnajärje muutumisel. (Estonian) [On the role of the information structure in the change of the Estonian word order]." *Keel ja Kirjandus* (*Language and Literature*) 11, 875–88.

Lippus, Pärtel, Tuuli Tuisk, Nele Salveste, and Pire Teras. Feb 2016. *Phonetic Corpus of Estonian Spontaneous Speech v.1.0.0*. http://hdl.handle.net/11297/1-00-0000-0000-0000-0003-1.

Jacob A. Long. 2017. *jtools: Analysis and Presentation of Social Scientific Data*. R package version 0.9.3. https://cran.r-project.org/package=jtools.

Tim Mahrt. October 2016. *Lmeds: Language markup and experimental design software*. Computer Program. https://github.com/timmahrt/LMEDS.

McCauley Stewart M. and Morten H. Christiansen. 2015. "Individual differences in chunking ability predict on-line sentence processing." In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. p. 1553–8. Austin, TX: Cognitive Science Society.

McQueen James M. and Taehong Cho. 2003. "The use of domain-initial strengthening in segmentation of continuous English speech." In: *Proceedings of the 15th International Congress of Phonetic Sciences*. p. 2993–6. Adelaide: Causal Productions.

Männel, Claudia and Angela D. Friederici. 2016. "Neural correlates of prosodic boundary perception in German preschoolers: If pause is present, pitch can go." *Brain Research* 1632, 27–33. http://www.sciencedirect.com/science/article/pii/S0006899315009488.

Müürisep, Kaili and Helen Nigol. 2008. *Where do parsing errors come from: the case of spoken estonian*. p. 161–8. Berlin, Heidelberg: Springer-Verlag.

Nakai, Satsuki, Sari Kunnari, Alice Turk, Kari Suomi, and Riikka Ylitalo. 2009. "Utterance-final lengthening and quantity in northern Finnish." *Journal of Phonetics* 37(1), 29–45. http://www.sciencedirect.com/science/article/pii/S0095447008000429.

Nash, John C. 2014. "On best practice optimization methods in R." *Journal of Statistical Software* 60(2), 1–14. https://www.jstatsoft.org/v60/i02/.

Nash, John C. and Ravi Varadhan. 2011. "Unifying optimization algorithms to aid software system users: optimx for R." *Journal of Statistical Software* 43(9), 14. https://www.jstatsoft.org/v43/i09/.

Nespor, Marina and Irene Vogel. 1986. *Prosodic phonology*. Dordrecht: Foris.

Niu, Ruochen and Timothy Osborne. 2019. "Chunks are components: A dependency grammar approach to the syntactic structure of mandarin." *Lingua* 224, 60–83. http://www.sciencedirect.com/science/article/pii/S0024384118308350.

Oller, D. Kimbrough. 1973. "The effect of position in utterance on speech segment duration in English." *The Journal of the Acoustical Society of America* 54(5), 1235–47. doi: https://doi.org/10.1121/1.1914393.

Ordin, Mikhail, Leona Polyanskaya, Itziar Laka, and Marina Nespor. 2017. "Cross-linguistic differences in the use of durational cues for the segmentation of a novel language." *Memory & Cognition* 45(5), 863–76. https://doi.org/10.3758/s13421-017-0700-9.

O'Shaughnessy, Douglas. 1979. "Linguistic features in fundamental frequency patterns." *Journal of Phonetics* 7(2), 119–45. https://www.sciencedirect.com/science/article/pii/S0095447019310459.

Peters, Benno. 1999. "Prototypische intonationsmuster in deutscher lese- und spontansprache." In: *Arbeitsberichte* (*AIPUK*), edited by Kohler, Klaus J., Vol. 34. p. 1–175. Kiel: IPDS.

Peters, Benno. 2005. "Weiterführende untersuchungen zu prosodischen grenzen indeutscher spontansprache. [ingl. further studies on prosodic boundaries in German spontaneous speech.]." In: *Arbeitsberichte* (*AIPUK*), edited by J. Kohler, K., Kleber, F., PetersVol. 35a. p. 203–345. Kiel: IPDS.

Peters, Benno, Klaus J. Kohler, and Thomas Wesener. 2005. "Phonetische merkmale prosodischer phrasierungin deutscher spontansprache [engl. prosodic structures in German spontaneous speech]." In: *Arbeitsberichte* (*AIPUK*), edited by Kohler, Klaus J., Felicitas Kleber, Benno Peters, Vol. 35a. p. 143–84. Kiel: IPSD.

Petrone, Caterina, Hubert Truckenbrodt, Caroline Wellmann, Julia Holzgrefe-Lang, Isabell Wartenburger, and Barbara Höhle. 2017. "Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists." *Journal of Phonetics* 61, 71–92. http://www.sciencedirect.com/science/article/pii/S0095447017300049.

Pierrehumbert, Janet B. 1980. *The phonology and phonetics of English intonation*. Ph.D. thesis, The Massachusetts Institute of Technology.

de Pijper, Jan Roelof and Angelien A. Sanderman. 1994. "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues." *The Journal of the Acoustical Society of America* 96(4), 2037–47. doi: https://doi.org/10.1121/1.410145.

Price, Patti, Mari Ostendorf, Stefanie Shattuck-Hufnagel, and Cynthia Fong. 1991. "The use of prosody in syntactic disambiguation." *Journal of the Acoustical Society of America* 90, 2956–70.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Austria: R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Raudvere, Uku and Kristel Uiboaed. 2018. *Uuema eesti ilukirjanduse mitmikute loendid*. doi: http://dx.doi.org/10.15155/re-8; http://datadoi.ee/handle/33/41.

Reichel, Uwe. 2011. "The CoPaSul intonation model." In: *Elektronische Sprachverarbeitung*, edited by Kroeger, B. and P. Birkholz, p. 341–8. Dresden: TUD Press.

Riesberg, Sonja, Janina Kalbertodt, Stefan Baumann, and Nikolaus P. Himmelmann. 2020. "Using rapid prosody transcription to probe littleknown prosodic systems: The case of papuan malay." *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1), 1–35.

Sanders, Lisa D. and Helen J. Neville. Dec. 2000. "Lexical, syntactic, and stress-pattern cues for speech segmentation." *Journal of speech, language, and hearing research: JSLHR* 43(11193954), 1301–21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2572147/.

Schafer, Amy Jean. 1997. *Prosodic parsing: The role of prosody in sentence comprehension*. Ph.D. thesis, doctoral Dissertations Available from Proquest. AAI9809396.

Schafer, Amy J., Shari R. Speer, and Paul Warren. 2005. *Prosodic influences on the production andcomprehension of syntactic ambiguity in a game-based conversation task*, p. 209–25. Cambridge: MIT Press.

Schafer, Amy Jean, Shari R. Speer, Paul Warren, and S. White. 2000. "Intonational disambiguation in sentence production and comprehension." *Journal of Psycholinguistic Research* 29(2), 169–82. doi: https://doi.org/10.1023/A:1005192911512.

Schegloff, Emanuel A. 1996. "Turn organization: one intersection of grammar and interaction." In: *Interaction and Grammar*, edited by Ochs, Elinor, Emanuel A. Schegloff, Sandra A Thompson, p. 52–133. Cambridge: Cambridge University Press.

Selkirk, Elisabeth O. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Shattuck-Hufnagel, Stefanie and Alice E. Turk. 1996. "A prosody tutorial for investigators of auditory sentence processing." *Journal of Psycholinguistic Research* 25(2), 193–247. doi: https://doi.org/10.1007/BF01708572.

Simon, Anne Catherine and George Christodoulides. 2016. "Perception of prosodic boundaries by naïve listeners in French." In: *Speech Prosody 2016, 31 May–3 June 2106*, Boston, USA. p. 1158–62. Lous Tourils (France): The International Speech Communication Association (ISCA).

Snedeker, J. and John C. Trueswell. 2003. "Using prosody to avoid ambiguity: Effects of speaker awareness and referential context." *Journal of Memory and Language* 48(1), 103–30.

Steinhauer, Karsten, Kai Alter, and Angela D. Friederici. Feb 1999. "Brain potentials indicate immediate use of prosodic cues in natural speech processing." *Nature Neuroscience* 2, 191–6.

Strangert, Eva. 1997. "Relating prosody to syntax: Boundary signaling in Swedish." In: *Proceedings of the 5th European Conference on Speech Communication and Technology*. edited by Kokkinakis, G., N. Fakotakis, E. Dermatas, p. 239–42. Lous Tourils (France): The International Speech Communication Association (ISCA).

Streeter, Lynn A. Dec 1978. "Acoustic determinants of phrase boundary perception." *The Journal of the Acoustical Society of America* 64(6), 1582–92.

Swerts, Marc, Don G. Bouwhuis, and René Collier. 1994. "Melodic cues to the perceived 'finality' of utterances." *The Journal of the Acoustical Society of America* 96(4), 2064–75, doi: https://doi.org/10.1121/1.410148.

Tael, Kaja. 1988. *Sõnajärjemallid eesti keeles* (*võrrelduna soome keelega*). (*Estonian*) [*Word order patterns in Estonian in comparison with Finnish.*] Technical Report, Tallinn: Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Thorsen, Nina Gro/nnum. 1985. "Intonation and text in standard Danish." *The Journal of the Acoustical Society of America* 77(3), 1205–16, doi: https://doi.org/10.1121/1.392187.

Trouvain, Jürgen, William Barry, Claus Nielsen, and Ove Kjeld Andersen. 1998. "Implications of energy declination for speech synthesis." In: *Speech Synthesis: Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, Australia, November 1998*, edited by Edginton, Mike, Implications of Energy Declination for Speech Synthesis; Conference date: 19-05-2010. p. 47–52.

Truckenbrodt, Hubert. 1999. "On the relation between syntactic phrases and phonological phrases." *Linguistic Inquiry* 30(2), 219–55.

Turk, Alice E. and Stefanie Shattuck-Hufnagel. Oct. 2007. "Multiple targets of phrase-final lengthening in American English words." *Journal of Phonetics* 35(4), 445–72.

Ulbrich, Christiane. 2002. "A comparative study of intonation in three standard varieties of German." *Proceedings of Speech Prosody* 2002, 671–4.

Vilkuna, Maria. 1995. "Discourse configurationality in Finnish." In: *Discourse configurational languages*, edited by Kiss, K .E., p. 244–68. New York: Oxford University Press.

Wagner, Michael and Michael McAuliffe. 2019. "The effect of focus prominence on phrasing." *Journal of Phonetics* 77, 100930.

Wagner, Michael and Duane G. Watson. Jan. 2010. "Experimental and theoretical advances in prosody: A review." *Language and cognitive processes* 25(22096264), 905–45. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3216045/.

Watson, Duane and Edward Gibson. 2005. "Intonational phrasing and constituency in language production and comprehension." *Studia Linguistica* 59(2/3), 279–300.

Wheeldon, Linda, Natalie Ohlson, Aimee Ashby, and Sophie Gator. Aug 2013. "Lexical availability and grammatical encoding scope during spoken sentence production." *Quarterly Journal of Experimental Psychology* (*2006*) 66(8), 1653–73.

Wheeldon, Linda and Mark Smith. 2003. "Phrase structure priming: A short-lived effect." *Language and Cognitive Processes* 18(4), 431–42. doi: https://doi.org/10.1080/01690960244000063.

White, Laurence, Silvia Benavides-Varela, and Katalin Mády. 2020. "Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues?." *Journal of Phonetics* 81, 100982. https://www.sciencedirect.com/science/article/pii/S0095447020300735.

Wightman, Colin W., Stefanie Shattuck-Hufnagel, Mari Ostendorf, and Patti J. Price. 1992. "Segmental durations in the vicinity of prosodic phrase boundaries." *The Journal of the Acoustical Society of America* 91(3), 1707–17. doi: https://doi.org/10.1121/1.402450.

Winter, B. 2019. *Statistics for Linguists: An Introduction Using R*, 1st ed. New York, London: Routledge.

Yang, Xiaohong, Xiangrong Shen, Weijun Li, and Yufang Yang. 2014. "How listeners weight acoustic cues to intonational phrase boundaries." *PLOS ONE* 9(25019156), e102166. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4096911/.