Laboratory Phonology
Journal of the Association for Laboratory Phonology

Ots, N. (2024). Planning sentences and sentence intonation in Estonian. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 15(1), pp. 1–50. DOI: https://doi.org/10.16995/labphon.8722

Open Library of Humanities

# Planning sentences and sentence intonation in Estonian

**Nele Ots,** Department of Linguistics, Goethe University, Frankfurt am Main, Germany, ots@em.uni-frankfurt.de

The notion of advance planning of sentence intonation is grounded in the positive correlation between the sentence-initial intonation peaks and sentence duration. This study examined real-time sentence planning and intonation using visual world speech production. In two eye-tracking experiments, native Estonian speakers described transitive events involving multiple actors. Conceptual complexity of the resulting picture descriptions was manipulated through a pictorial design, while sentence length was controlled for by manipulating specific task characteristics. In Experiment I, conceptual complexity of the picture descriptions varied together with linguistic complexity, while linguistic complexity was held constant in Experiment II. As the conceptual complexity of utterances increased, the duration of naming gazes also increased, indicating less incremental conceptual planning. Notably, while utterance-initial intonation peaks did not correlate with the relative duration of naming gazes, they were influenced by utterance length. These findings highlight advance planning of intonation in Estonian. Furthermore, they suggest that intonation planning depends on linguistic information that is rapidly activated after establishing a comprehensive conceptual framework during earliest stages of preverbal planning.

# 1 Introduction

For the utterance of their ideas, speakers usually need to find the right words (lexical retrieval) and an appropriate ordering of these words (syntactic encoding). In spoken language, not just the wording itself matters; the way in which the words are said is also important. A large part of how words are spoken is accounted for by the notion of sentence prosody. Sentence prosody refers to relative changes in speech melody (i.e., intonation) and rhythm in language chunks larger than a word (e.g., a syntactic phrase, clause or intonation phrase) or in single words, in cases where they constitute single-word utterances.

A number of psycholinguistic models account for how speakers transform their ideas into grammatical utterances, that is, the planning of sentences (for an overview, see Meyer, Roelofs, & Brehm, 2019). However, accounts of finding the appropriate intonation contours are developing (Fromkin, 1971; Garrett, 1975, 1980; Keating & Shattuck-Hufnagel, 2002; Shattuck-Hufnagel, 2019; Wheeldon & Lahiri, 1997; Wheeldon & Smith, 2003), and it is still unclear how the production of sentence prosody pertains to language-related cognitive processes in sentence planning.

## 1.1 Sentence planning

Theories of sentence planning investigate the cognitive processes that lead to grammatical utterances. The prevailing psycholinguistic models of sentence planning propose that utterances are planned by switching between the stages of message generation, grammatical encoding and phonological encoding (Garrett, 1975, 1980; Levelt, 1989; Levelt et al., 1991). Specifically, spoken sentences are planned by first converting messages (which are the output of conceptualization) into abstract lexical representations (i.e., lemmas), then assigning the syntactic functions (i.e., the process of grammatical encoding or linguistic formulation) and finally compiling the words' base forms and affixes (i.e., the process of phonological encoding). The three stages of planning are not necessarily completed prior to the start of articulation. For example, articulation of the first word of an utterance can be accompanied by simultaneous grammatical encoding of subsequent parts of the utterance (Levelt, Roelofs, & Meyer, 1999; Levelt et al., 1991). Thus, among a number of other aspects of speaking, utterances involve not only articulation but also concurrent conceptual planning, i.e., message generation, lemma retrieval, and grammatical and phonological encoding.

One of the main concerns in the study of sentence planning is how far in advance speakers plan their utterances before they start articulating. Speakers can begin speaking as soon as the linguistic form of the initial part of a message is activated, and plan further bits and pieces of their utterances whilst they are already speaking (Gleitman, January, Nappa, & Trueswell, 2007; Levelt, 1989; Myachykov & Tomlin, 2008; Tomlin, 1995). Alternatively, speakers might also plan further in advance, having the phonological structure of the entire utterance ready just before articulation (Ferreira & Swets, 2002; Oppermann, Jescheniak, & Schriefers, 2010). In language production, the scope of planning and the degree of incrementality are closely related. A smaller planning scope leads to a more incremental planning strategy. The more incremental planning strategy requires less working memory (Cole & Reitter, 2019; Slevc, 2011) and therefore, it is cognitively more efficient. However, the cognitive efficiency does not necessarily ensure the fluency of spontaneous speech production. Speech errors are more likely when smaller working memory span is engaged in speech production (Badecker & Kuminiak, 2007; Hartsuiker & Barkhuysen, 2006; Slevc, 2011).

The theory of hierarchical incrementality accounts for the fluency of spontaneous speech production by proposing a large scope for planning messages and scopes of varying sizes for the grammatical and phonological encoding (Bock, Irwin, & Davidson, 2004; Griffin & Bock, 2000; Kuchinsky, Bock, & Irwin, 2011; Wheeldon, Ohlson, Ashby, & Gator, 2013). In other words, planning messages proceeds non-incrementally while grammatical and phonological encoding can vary, proceeding less or more incrementally. According to the account of hierarchical incrementality, the stage of message generation outputs a rudimentary conceptual framework (e.g., who is doing what to whom), which then guides lemma retrieval, assembly of grammatical structures and phonological encoding (Bock et al., 2004; Griffin & Bock, 2000; Kuchinsky et al., 2011; Wheeldon et al., 2013). In other words, speakers may start an utterance without having the right words for the subsequent content of their utterance, but the difference is that they have a broad idea about the information they are about to articulate, piece by piece.

The majority of existing research suggests that speakers can effectively shift between highly incremental and less incremental strategies (for highly incremental planning, see Gleitman et al. 2007; Myachykov and Tomlin 2008; Tomlin 1995; for less incremental planning, see Bock et al. 2004; Griffin and Bock 2000; Kuchinsky et al. 2011; Wheeldon et al. 2013). Additionally, in an eye-tracking study, Konopka and Meyer (2014) found that message planning can also proceed more or less incrementally, depending on the linguistic characteristics of incipient utterances. Specifically, when the names of the first-mentioned actors/objects were relatively easy to retrieve, speakers initiated the utterance of the first word soon after gazing at the corresponding referent and proceeded with conceptualization and linguistic encoding for the rest of the utterance while already speaking. When the first-mentioned actors/objects could be named with a range of different terms (e.g., bunny, rabbit, hare), the gaze was more frequently directed towards one and then the other referent before the start of the utterance. In other words, the scope of message planning varied as a function of lexical diversity. Similarly, some cross-linguistic studies suggest that the extent of how far in advance the message is conceptualized before speaking depends on the type of language, with morphological case-marking languages employing the least incremental planning strategies (Norcliffe, Konopka, Brown, & Levinson, 2015; Sauppe, Norcliffe, Konopka, Valin, & Levinson, 2013). Thus, speakers can choose either incremental or less incremental message planning, depending on the characteristics of incipient utterances and the grammatical preferences of a particular language. In other words, they can either generate a rudimentary conceptual framework corresponding to the entire incipient utterance or proceed with linguistic encoding right after activating the first-mentioned concept.

Similar to these previous studies of message planning, the current study employs the eye-tracking technique to investigate the incrementality of conceptual planning, using visual scenes with varying complexity.

## 1.2 Intonation planning

As far as the relationship between planning sentences and sentence intonation is concerned, the relevant question is: at which point in the utterance planning does sentence prosody (i.e., prosodic phrasing and prominence) emerge? In other words, does the planning of intonation depend on the activation of conceptual, syntactic or phonological information of incipient utterances?

One of the earliest psycholinguistic proposals is that intonation contours, together with prosodic prominence, are decided right after lexical retrieval but before the assignment of syntactic functions (Fromkin, 1971). This conclusion was substantiated by the observation that intonation contours in English remained the same when speech errors involved the substitution of phonemes. Keating and Shattuck-Hufnagel (2002) argue that a number of phonological processes (e.g., strengthening, accentual lengthening and final lengthening in a variety of languages and stress clash in English) are informed by an independent abstract representation of sentence prosody which needs to be generated right after the grammatical but before the phonological encoding (see also Keating, 2006; Shattuck-Hufnagel 2019). For example, consonants that occur at phrase and clause boundaries are articulated more strongly than those occurring phrase-internally. According to Keating and Shattuck-Hufnagel (2002), the strengthening of sounds at the phrase boundaries is contingent on the access of clausal organization of upcoming speech. Thus, intonation planning is proposed to occur at the interface of grammatical and phonological encoding and to involve the generation of an abstract prosodic structure which, in turn, guides the retrieval of phonological syllable codes (Fromkin 1971; Keating, 2006; Keating & Shattuck-Hufnagel, 2002; Shattuck-Hufnagel, 2019). In other words, intonational phrasing and prominences are proposed to arise structurally based on an abstract prosodic structure that refers to syntactic information activated at the level of grammatical encoding (Keating & Shattuck-Hufnagel, 2002).

Alternatively, sentence prosody might emerge incrementally during the process of phonological encoding (see, e.g. Levelt, 1989; Levelt et al., 1999). The prosodic planning processes in this model feed the speech production system with intonational parameters in parallel to the incremental retrieval of the phonological codes of syllables. For the intonational parameters, the model considers the declination trend, key and register, and pre-nuclear and nuclear tones (see Levelt, 1989, 398–405). Only the parameters of register and key are set globally ahead of phonological retrieval; all other parameters can be set incrementally with almost no preview of upcoming linguistic material. In Levelt's terms, the register and key seem to refer to low and high levels of voice pitch. For each utterance, the speech production system chooses some global values for the key (phonetically topline F0) and register (baseline F0) and then it models each F0 movement (acoustic approximation of a pitch accent) incrementally as a deviation (of the selected key size) of register. The decision to produce a prominence-lending nuclear pitch accent can be made with a minimum of one-word lookahead. Levelt (1989) admits that some types of pre-nuclear pitch accents might require a somewhat longer preview of an incipient utterance, but these are rather infrequent in casual speech. According to the model in Levelt (1989), a prosodic break can also be decided with no more lookahead than one word. Thus, the prosodic structure with prominences and intonational phrasing is proposed to emerge incrementally, syllable by syllable or foot by foot (a stressed syllable with preceding or subsequent unstressed syllables constitute a prosodic foot). Put differently, sentence intonation is not planned at all and it evolves incrementally as the phonological structure is retrieved (Levelt, 1989).

Relatedly, the study of the syntax–phonology interface is very much concerned with the relationship between the syntactic structure and the prosodic structure. By definition, the accounts of the syntax–phonology interface do not deal with planning sentences and sentence intonation, but for the current purposes, they might provide some further insights into how syntactic and prosodic components of language might merge into an utterance during sentence planning. Most of these theories converge on the notion of the so-called structural planning of sentence intonation (Keating & Shattuck-Hufnagel, 2002). These accounts agree that a distinctive prosodic structure is generated after the completion of morphosyntactic movement operations (Elordieta & Selkirk, 2022; Kratzer & Selkirk, 2020; Lee & Selkirk, 2022; Nespor & Vogel, 1986; Selkirk, 1980, 1981, 1986, 2011). The accounts disagree on which type of morphosyntactic information the generation of prosodic structure requires. For instance, the prosodic structure may be generated based on the boundaries of major syntactic constituents (so-called edge-based theories; see, e.g. Selkirk, 1986), on phonological constraints that map the phonological constituents to syntactic constituents, (so-called Match Theory; see, e.g., Selkirk, 2011) or on morphophonological information (Elordieta & Selkirk, 2022; Kratzer & Selkirk, 2020; Lee & Selkirk, 2022). The less pronounced idea of the syntax–phonology interface states that the prosodic structure does not exist and that there are no prosodic constituents or domains where phonological processes may apply (Kaisse, 1985; Odden, 1987, 1990). This idea might conform with the notion of highly incremental generation of sentence prosody in Levelt (1989).

Most recently, Himmelmann (2022) has proposed a contrasting idea that diachronically, syntactic phrases might have developed from prosodic chunks. Synchronically, the proposal distinguishes between chunks of speech where prosodic scaffolding does not interrupt the syntactic cohesion (e.g., *part of the [uhm] route*) and chunks where the prosodic phrasing and prominence change the function of the chunk (e.g., detached constructions such as topics, afterthoughts, parentheses, appositions, quotes). With the help of this distinction, Himmelmann (2022) argues for the independence of the prosodic structure from the syntactic structure and posits an adjacency principle to argue that the new types of syntactic phrases emerge from prosodic phrasing. This idea implies, at least for some types of syntactic structures, that the grammatical encoding processes refer to prosodic components of language. The notion of prosody emerging before syntactic structure suggests that the processing of grammar for comprehension and speaking is highly flexible. Speakers can rely on existing and prosodically robust phrase structures or create new structures at any time.

A phonetic notion of the advance planning of sentence intonation offers additional insights into the matter of intonation planning. The advance planning of sentence intonation is grounded in the observation that intonation peaks at the very beginning of intonation phrases tend to be higher in longer sentences rather than shorter ones (see e.g., Cooper & Sorensen, 1981; Liberman & Pierrehumbert, 1984; Yuan & Liberman, 2014). This finding suggests that speakers anticipate the length of an incipient intonation phrase (which sometimes corresponds with a sentence or a full clause) while producing the first pitch-accented word. Relatedly, a number of studies observe that the low Fundamental Frequency (F0; the acoustic index of pitch and a proxy for sentence intonation) at the ends of utterances is usually quite invariant in a given speaker (Cooper & Sorensen, 1981; Liberman & Pierrehumbert, 1984; Yuan & Liberman, 2014), indicating the limitations of manipulating the lower register of voice pitch (Liberman & Pierrehumbert, 1984). Therefore, speakers may try to avoid the limitations of lower pitch register and control for the height of sentence-initial intonation peaks and the overall decline of intonation contours.

Indeed, a number of studies show that linguistically driven fluctuations of F0 correlate only weakly with subglottal air pressure (see e.g., Atkinson, 1978; Fuchs, Petrone, Krivokapić, & Hoole, 2013; Honda, 2004; Honda, Hirai, Masaki, & Shimada, 1999; Strik & Boves, 1992), as they are controlled with the help of laryngeal muscles (see e.g., Honda, 2004). Thus, speakers are indeed able to intentionally control the height of intonation peaks and avoid a too-deep drop of voice pitch towards the ends of utterances.

Moreover, the advance planning of sentence intonation has been documented for a variety of languages (for Estonian, see Asu, Lippus, Salveste, and Sahkai 2016; for English, see Cooper and Sorensen 1981; Liberman and Pierrehumbert 1984; Yuan and Liberman 2014; for Danish, see Thorsen 1980; for Mandarin Chinese, see Yuan and Liberman 2014). Therefore, the length-dependent raising of utterance-initial intonation peaks can be taken to reliably index the advance planning of sentence intonation (e.g., Cooper & Sorensen, 1981; Liberman & Pierrehumbert, 1984; Prieto, D'Imperio, Elordieta, Frota, & Vigário, 2006; Yuan & Liberman, 2014). An intriguing question is at which stage of sentence planning utterance-initial intonation peaks are scaled, and how does the concept of advance planning of sentence intonation interact with the cognitive model of speaking?

Most of the ideas found in the literature on sentence planning and the syntax–phonology interface suggest that intonation planning interfaces with grammatical and phonological encoding of messages and involves the generation of an abstract prosodic structure that controls for the phonological phenomena in spoken sentences. The length-dependent scaling of sentence-initial intonation peaks, however, entails an early selection of the intonation parameters for the verbalization of thoughts. Therefore, intonation planning (at least for some tonal parameters) might also occur earlier, such as at the stage of message planning.

## 1.3 The current study

The length-dependent raising of intonation peaks at the beginning of utterances entails that speakers must be able to predict the ends of incipient utterances relatively accurately. Concerning the idea of highly incremental planning of spoken utterances as opposed to the hierarchically incremental account of sentence planning, the highly incremental planning cannot accommodate the existing evidence on the advance planning of sentence intonation. Specifically, how would speakers know to raise the utterance-initial intonation peaks when they are not sure about what else they will add to the first words of their utterances? Thus, an alternative account, namely the account of hierarchical incrementality, is considered to explain the underlying cognitive mechanisms of length-dependent raising of intonation peaks at the beginning of spoken utterances. On the assumption of the hierarchically incremental account of sentence planning, the generation of messages is more likely to proceed in larger rather than smaller planning increments, that is, less incrementally (for the varying degrees of incrementality in planning the conceptual component, see Konopka and Meyer 2014). Consequently, the conceptual representation of incipient utterances is also more likely to underlie the speakers' ability to accurately predict the ends of their utterances and to plan the declination trend of the F0 accordingly. Therefore, the current study proposes and tests the idea that the advance planning of sentence intonation may be related to conceptualization rather than grammatical or phonological encoding processes. In particular, the aim is to explore the relationship between the incrementality of conceptual planning and the tonal scaling of utterance-initial intonation peaks.

As discussed earlier, research has shown that the linguistic structure of emerging utterances can influence the incrementality of conceptual planning (Konopka & Meyer, 2014; Sauppe, 2017). For example, when referring to agents becomes more challenging (due to lexical diversity), the likelihood for larger increments of message planning increases (Konopka & Meyer, 2014). Additionally, Sauppe (2017) has demonstrated for German that the incrementality of conceptual planning depends on syntactic structure (active vs. passive sentences) and the position of the verb (second position vs. sentence-final). These studies, however, do not clarify the extent to which the conceptual complexity embedded within these linguistic factors contributes to the incrementality of conceptual planning. Therefore, the first goal of the study is to isolate the conceptual factors of message generation from the linguistic ones.

The second goal of the study is to delve into the relationship between intonation and sentence planning by investigating whether and to what degree the height of sentence-initial intonation peaks depends on the degree of incrementality of conceptual planning. A correspondence between sentence-initial intonation peaks and the size of the increments of conceptual planning is expected to demonstrate that the advance planning of sentence intonation relies on the conceptual planning processes. Alternatively, a positive correlation between sentence-initial intonation peaks and the length of utterances irrespective of the incrementality of conceptual planning would demonstrate that the advance planning of sentence intonation is contingent on the linguistic representation of upcoming speech.

The current study draws on the advance planning of Estonian sentence intonation. Estonian is consistently characterized by its mainly declining intonation contours (Asu, 2004). Utterances end in a low register or, in terms of autosegmental-metrical phonology, with a low boundary tone (L%; Asu 2005). Content words are most often characterized by falling F0 contours (H* + L; Asu and Nolan 1999). Thus, in comparison to fully–fledged intonation languages like English or German, the Estonian intonational inventory is considerably less variable. A recent corpus study Asu et al. (2016) found that the duration of prosodic chunks (defined as inter-pausal units) contribute to the height of the phrase-initial intonation peaks in spontaneous speech such that the longer a prosodic chunk, the likelier a higher phrase-initial intonation peak. Given this finding, the tonal uniformity of Estonian intonation, and the ability to measure F0 in comparable tonal contexts, Estonian presents an excellent test bed for the examination of the relationship between the incrementality of conceptual planning and the advance planning of sentence intonation.

## 2. Method

Two speech production experiments investigated how different degrees of incrementality in conceptual planning and variations in sentence length affect the advance planning of sentence intonation. Native speakers of Estonian were asked to describe pictures of simple interactions showing either two (agent and patient), three (agent, patient and an attribute/distractor) or four (agent, patient, and attribute plus the distractor) actors/objects by mentioning the event and interacting actors – the agent and patient. As a result, the majority of utterances constituted transitive sentences with a subject-verb-object ordering (SVO) in which the agent was mentioned before the patient. Across the two experiments, the number of pictured actors and objects was manipulated to control the incrementality of conceptual planning such that the larger number of actors and objects in the pictures was hypothesized to initiate the generation of a more comprehensive rather than a limited relational framework (i.e., a larger planning increment of a message). Depending on the requirement to mention the attributes of the actors or not, the length of the picture descriptions varied within Experiment I and across the two experiments such that the speakers produced short and long sentences in Experiment I (e.g., *Mees sikutab eeslit* 'A man is pulling the donkey'; *Mees sikutab vana eeslit* 'A man is pulling the old donkey'; *Mees sikutab korviga eeslit* 'A man is pulling the donkey with a basket') and only short sentences in Experiment II (e.g., *Mees sikutab eeslit* 'A man is pulling the donkey').

Importantly, the length of utterances corresponded to the duration of utterances (in milliseconds) and the increase in length was bound to both syntactic and phonological factors (i.e., a greater number of words and morpho-syntactic complexity). Therefore, the variation of duration (i.e., the varying number of words in Experiment I and the same number of words in Experiment II) appoximated the linguistic encoding processes, encompassing grammatical and phonological encoding. The varying number of pictured actors and objects, i.e., the size of the visual array, estimated the conceptual planning. In other words, the different types of complexities served as a diagnostic for the conceptual and linguistic planning processes, and helped to isolate the stage of message planning from the processes of grammatical and phonological encoding.

To capture the degrees of incrementality in conceptual processing, the current study employs the method of recording speakers' eye movements before and whilst they describe pictures of simple events. As the naming of actors and objects is tightly coupled with the direction of the gaze in a visual display, the eye movements provide a real-time approximation of on-going encoding processes (see Griffin & Bock, 2000; Levelt et al., 1991). In particular, numerous studies have established that the stages of message generation and phonological encoding are clearly observable in the direction and duration of the gaze (Bock et al., 2004; Griffin, 2001; Griffin & Bock, 2000; Konopka & Meyer, 2014; Kuchinsky et al., 2011; Norcliffe et al., 2015; Sauppe, 2017; Sauppe et al., 2013). More specifically, speakers tend to look longer at the objects they are going to mention than the objects they do not need to mention (see, e.g., Levelt et al., 1991). Research has established that these longer viewing times are also present during the production of full sentences, but that they occur no earlier than around 1000 ms from the onset of a picture. For about 400 or 600 ms after picture presentation, speakers distribute their visual attention between the regions of different actors and objects involved in the depicted events (see, e.g., Griffin & Bock, 2000; Sauppe, 2017). Right before onset of speech (at about 1000 ms after picture presentation), the eyes are usually directed towards the actor/object that speakers mention first in their utterances and stay there longer, possibly until the onset of speech. In the following analysis (see Section 2.5), the long gazes at the first-mentioned actors/objects serve as landmarks of cognitive processing between the picture onset and the onset of speech. Around the onset of speech (at around 1800–2000 ms), the gaze typically moves to the actor/object mentioned second. It has been argued that the brief time frame of sharing visual attention at the onset of a picture facilitates subsequent rapid processing of the visual information in the order it is mentioned (Bock et al., 2004; Griffin & Bock, 2000). In other words, the brief apprehension of event gist approximates the generation of an initial conceptual frame (i.e., who is doing what to whom; Griffin and Bock 2000), which then guides the eyes along the actors/objects in the visual displays in the order of mentioning.

Several studies provide evidence that a number of linguistic factors affect the manner in which the event gist is apprehended by speakers. For instance, the distribution of eye movements in the putative time window of conceptual processing has been shown to vary as a function of word order (Norcliffe et al., 2015; Sauppe, 2017), morphological case marking (Sauppe et al., 2013), syntactic priming (Konopka & Meyer, 2014; van de Velde, Meyer, & Konopka, 2014) and ease of encoding the event and the names of actors/objects (Konopka & Meyer, 2014). In particular, these studies have detected that the event apprehension may only concentrate on the actor mentioned first in an incipient utterance. This pattern of eye movements indicates that speakers begin speaking by mentioning the actor/object they see first and consider the other parties of an event while they are already speaking. In terms of incrementality, this demonstrates *incremental* conceptual planning. Alternatively, speakers may attend to all actors/objects irrespective of first mention. This pattern of eye movements indicates that speakers first take time to look at all the actors/objects in the display and then start to speak after generating a rudimentary conceptual representation of all parties (and not of just one party) of an event. In other words, the conceptual planning proceeds *less or non-incrementally*. Therefore, the distribution of eye gazes in the earliest time frame of picture processing (200–1000 ms after picture onset) can be taken to approximate the incrementality of conceptual planning.

The combination of conceptual and linguistic factors in the current study resulted in visual displays with an increasing number of actors/objects. Earlier studies involving spoken responses have shown that the size of the visual array has a strong delaying effect on speech onset latencies (Elsner, Clarke, & Rohde, 2018; Gatt, Krahmer, van Deemter, & van Gompel, 2017). For larger visual arrays, the eye takes additional time to progress across the greater number of actors/objects. The resulting longer speech onset latencies may reflect the speakers' need to decide on the higher-level strategy for describing the content of more complex visual scenes (Elsner et al., 2018). In other words, larger visual arrays enforce less incremental conceptual planning.

The two different types of complexities (conceptual vs. linguistic) were predicted to impact the extent of incrementality in conceptual planning such that the degree of incrementality decreases with growing conceptual complexity and the length of picture descriptions. Larger increments of conceptual planning are expected to be reflected in a more even distribution of visual attention between interacting actors/objects. To establish a link between intonation planning and the conceptual planning stage, the size of conceptual planning increments (i.e., measure of eye movements) is predicted to align with the height of sentence-initial intonation peaks, irrespective of sentence length. Alternatively, if the intonation planning relates to linguistic encoding, the height of sentence-initial intonation peaks should vary based on utterance duration only (short vs. long sentences).

## 2.1 Stimuli and design

For visual stimuli (64 for Experiment I and 60 for Experiment II), pictures of simple actions were constructed such that they were expected to affect the incrementality of conceptualization and the length of utterances in the task of picture description. To manipulate the incrementality of conceptual planning, the number of animate as well as inanimate actors was set to increase across the four conditions. The size of the visual array was varied by crossing the two factors, Distractor and Separable Attribute, with two levels each. The factor Distractor accounted for the number of parties (two or three) involved in the depicted events. The events involving two parties contained only two interacting actors: an initiator of an action, the *agent*, and an undergoer of an action, the *patient* (see **Figures 1a, 1b, 1e** and **1f**). The events involving three parties included a third actor, the so-called distractor (see **Figures 1c, 1d, 1g** and **1h**). For Experiment I, the distractor was designed to resemble the patient involved in the interaction (see **Figures 1c** and **1d**). For Experiment II, the distractor belonged to the same semantic group as the interacting patient (see **Figures 1g** and **1h**). Nevertheless, the appearance of the distractor was designed to cue membership of a different subgroup within the given semantic group (e.g., goats and donkeys are different species of farm animals).

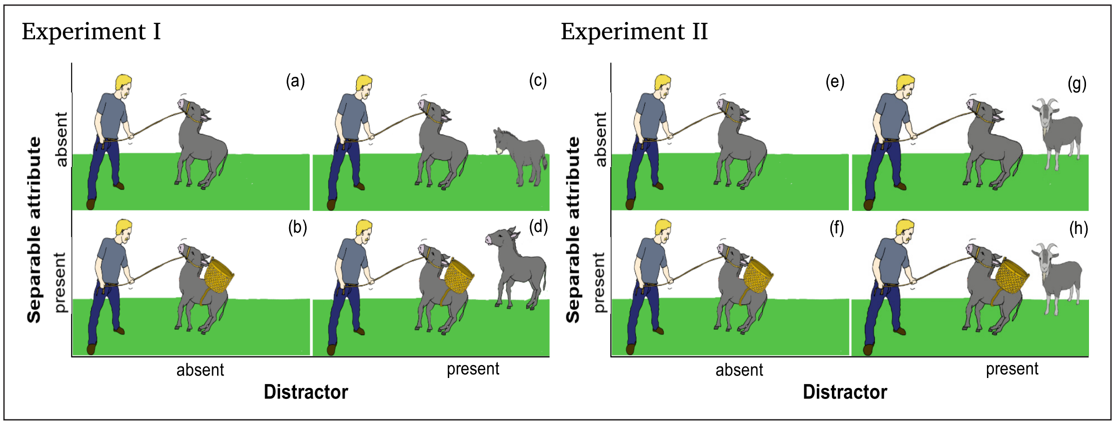

**Figure 1:** Designs of the two visual world speech production experiments. In Experiment I, speakers were asked to mention the two interacting characters. For the conditions with distractors present, they were additionally instructed to mention a feature of the target patient to identify this character, resulting in descriptions including noun phrases with attributive adjectives and nouns as noun phrase modifiers. In Experiment II, speakers were asked to identify the target patient with a distinguishing name (e.g., *a baby* instead of *a child* or *children*).

The factor Separable Attribute accounts for the differing features of the patients. The inseparable features such as colour and size occurred across both experiments and all four conditions (e.g., the purple colour of a car or the adulthood of a donkey). The separable features were objects or patterns that only occurred in some of the conditions (e.g., a surfboard on the top of a car or a basket on the back of a donkey; see **Figures 1b, 1d, 1f, 1h**). While the inseparable attributes did not necessarily result in an increase in the number of actors/objects in the visual array, the separable attributes did increase the number of actors/objects in the pictures. Therefore, the factor Separable Attribute is mainly responsible for the conceptual complexity of the picture descriptions.

The factor Distractor interacted with a concrete task setting such that in Experiment I, it triggered longer and shorter picture descriptions. Specifically, speakers were asked to use more complex references for the interacting patient as soon as they detected a distractor in a picture. Thus, the presence of distractors made the attributes visible and relevant for the task of picture description. Speakers were requested to mention the inseparable attributes in the form of attributive adjectives (see Example 1a) and separable attributes in the form of attributive nouns (see Example 1b).

(1) a. suur-t eesli-t
big.*SG-PART* donkey.*SG-PART*
'big donkey'

b. korvi-ga eesli-t
basket.*SG-COM* donkey.*SG-PART*
'a donkey with a basket'

Thus, the factor Distractor ensured that the picture descriptions varied in linguistic complexity such that the sentences mentioning the interacting actors only were the shortest (see Example 2a), the sentences including attributive nouns (see Example 2c) constituted the longest utterances, and the sentences including attributive adjectives fell in between (see Example 2b).

(2) a. mees sikuta-b eesli-t
man.*SG.NOM* pull-*3SG* donkey.*SG-PART*
'a/the man is pulling a/the big donkey'

b. mees sikuta-b suur-t eesli-t
man.*SG.NOM* pull-*3SG* big.*SG-PART* donkey.*SG-PART*
'a/the man is pulling a/the big donkey'

c. mees sikuta-b korvi-ga eesli-t
man.*SG.NOM* pull-*3SG* basket.*SG-COM* donkey.*SG-PART*
'a/the man is pulling a/the donkey with a/the basket'

In Experiment II, the factor Distractor influenced the concreteness of reference. Speakers were instructed to refer to specific instances or examples within the given semantic group, for example by using the name “cat” rather than a general category label like “pet.” Thus, the resulting picture descriptions for all four conditions mentioned the two interacting actors without any further specifications, e.g., “the man is pulling a donkey.” In this way, combination of the factor Distractor and the specific task setting in Experiment II triggered short utterances where the linguistic complexity was held constant but the conceptual (visual) complexity still increased across the four conditions. Therefore, the factor Distractor is taken to account for the linguistic complexity across both experiments. However, it is worth noting that while controlling for the linguistic complexity of the picture descriptions, the factor Distractor also contributed to a greater number of actors/objects in the visual display and augmented conceptual complexity of utterances.

The stepwise inclusion of attributes and distractors across the four conditions of the two experiments ensured a gradual increase in the conceptual complexity of the messages and allowed for the disengagement of the conceptual factors (i.e., Separable Attribute) from the linguistic factors of sentence planning (i.e., Distractor). In particular, the pictures including the two interacting actors together with the separable attribute (see **Figures 1b** and **1f**) are conceptually more complex than the pictures of the two actors only (see **Figures 1a** and **1e**). The conceptual complexity of the pictures including the two actors and the distractor (**Figures 1c** and **1g**) is comparable with the conceptual complexity of pictures including the two actors and the separable attribute (**Figures 1b** and **1f**) because they both include three referrable actors/objects. The pictures, including the two actors, the attribute and the distractor (**Figures 1d** and **1h**), present the most conceptually complex conditions.

The target pictures (64 for Experiment I and 60 for Experiment II) in the four conditions, mirror-reversed to counterbalance the leftward and rightward placement of agents and patients, were distributed across eight lists according to a Latin Square design. Given the Latin Square design, each participant saw 16 different pictures for each of the four conditions in Experiment I and 15 different pictures for each of the four conditions in Experiment II. In Experiment I, another eight lists reversed for the order of target trials were created to control for the effects of fatigue. Within a list, the target pictures were interspersed among 106 filler pictures in Experiment I and 105 filler pictures in Experiment II. The filler pictures depicted a range of events with one, two or multiple parties. While the target pictures were intended to be described with transitive sentences, the filler pictures could be described with a variety of structures. The pictures in the lists were ordered to avoid semantic overlap and repetition of content words between the two consecutive trials. Seven additional pictures served as examples and practice trials. Altogether, the 16 lists in Experiment I comprised 177 pictures and the 8 lists in Experiment II comprised 165 pictures.

## 2.2 Subjects

In total, 53 adult native speakers of Estonian participated voluntarily in Experiment I. For this study, the results of 45 speakers (32 females with a mean age of 31.3 years and 12 males with a mean age of 31.5 years) were included in the analysis. Of the excluded participants, five speakers started speaking very slowly (longer than 3000 ms) and/or appeared not to understand the task. One speaker reported being a native speaker of both Estonian and Finnish, and the Finnish influences were clearly audible in their Estonian speech production. Due to a recording error, no sound file was created for the two final speakers. For the included 45 participants, Estonian was the first native language, no other first languages were reported, and no data for second language experience was collected.

Seventy-four adult native speakers of Estonian participated for the incentive of 5 euros in Experiment II. For this study, the results of 68 adult native speakers of Estonian (56 females with a mean age of 26.4 years and 12 males with a mean age of 22.5 years) were included in the analysis. Of the exclusions, the eyes of two participants could not be tracked and the procedure was cancelled right after the failure to scale the eye tracker. Three of the speakers started speaking very slowly (longer than 3000 ms), one of which was strongly influenced by Russian (which was reported as a second native language). Due to a recording error, eye-tracking data is missing for the final speaker who was not included. For the 68 included participants, Estonian was the first native language. Only one speaker reported English as being her second native language, and no data for foreign languages was collected.

The included participants of both experiments had normal or corrected-to-normal vision. Across the two experiments, no language deficiencies were reported. The recruiting procedures made sure that the participants of Experiment II did not participate in Experiment I.

## 2.3 Procedure

Speakers were recorded at the University of Tartu. For Experiment I, pictures were presented on an SR Research EyeLink 1000 Plus tower-mounted eye tracker (sampling rate 1000 Hz, distance to participant ca. 70 cm, refresh rate of the display 60 Hz). Speakers were instructed to produce informative descriptions of the pictures by mentioning all interacting actors in action (e.g., *Mees sikutab eeslit* ‘the man is pulling a donkey’). They were asked to exclude the non-interacting actors from their descriptions but to take the presence of them into account by mentioning the distinguishing features of the interacting patients. Thus, the size of the visual array, together with the concrete task setting, elicited shorter and longer sentences varying in the degree of conceptual complexity. The inseparable and separable attributes in Experiment I were designed to encourage the use of adjective phrases (e.g., *paljas tita* ‘naked baby’; *lilla auto* ‘purple car’) and modifier noun phrases (e.g., *pudeliga tita* ‘a baby with a bottle’; *surfilauaga autot* ‘a car with a surfboard on top'). More specifically, participants of Experiment I were explicitly asked to mention the inseparable and separable attributes as soon as they saw a third non-interacting actor (i.e., the distractor) in the picture.

For Experiment II, the pictures were presented on an SR Research EyeLink 1000 Plus desktop-mounted eye tracker (sampling rate 1000 Hz, distance to participant ca. 90 cm, refresh rate of the display 60 Hz). Speakers were again instructed to produce informative sentences by describing the actions and mentioning the two interacting actors. For the conditions with distractors present, they were requested to find a more concrete name for the interacting patient (e.g., *tita* 'baby' instead of the category label *laps* 'child' or *lapsed* 'children') and they were explicitly asked not to use the noun phrase modifiers (e.g., small child) or plural nouns (e.g., children). Thus, the varying number of actors together with the task setting were designed to elicit short sentences varying in the degree of conceptual complexity. In addition, to ensure a better cross-linguistic comparability, the participants of Experiment II were asked to start speaking faster when they appeared to be slow in performing the task.

The target trials in both experiments were preceded with a training session containing seven practice trials. For the recording procedure, each trial was initiated with a fixation cross at the centre top of the screen. The fixation cross was followed by a black dot at the centre top of the screen. The experimenter made sure that the eye stayed on the black dot and started the presentation of the picture with a mouse click. This measure was necessary to start the sound recording simultaneously to picture presentation in the Experiment Builder and to ensure that the first fixation would not fall on the actors in the pictures. If speakers moved their eyes from the dot, they were asked to look at the dot again. After speakers had completed their descriptions, the experimenter initiated the next trial with another mouse click. The sum of the trial durations in Experiment I, with a list of 177 pictures, was about 20.1 minutes (SD = 3.21 minutes); and in Experiment II, with a list of 165 pictures, was about 18.01 minutes (SD = 3.53 minutes). Together with the scaling procedures and drift corrections between the trials, both experiments took about 40 to 60 minutes of the participants' time.

## 2.4 Preprocessing

Altogether, 2871 descriptions of target pictures were recorded in Experiment I, and 4020 descriptions of target pictures were recorded in Experiment II. Areas of Interest (AOIs) were defined independently of actual eye fixations to cover the agents, patients, attributes and distractors, and a margin of 100 pixels around them (Holmqvist et al., 2011). All fixations were scored as falling within the interest area of the agent, the patient, the attribute, the distractor or outside the designated AOIs (scored as empty). The AOIs of the inseparable attributes (e.g., colour, material, size) coincided with the AOIs of the patients, while the AOIs of the separable attributes were either partly overlapping with the patients or separated from them. The entire trial was disregarded when (i) the first fixation fell on the region of a defined AOI instead of the fixation dot, (ii) the first fixation on the region of an entity occurred later than 400 ms after picture onset, or (iii) the time delay between the two consecutive fixations was longer than 600 ms (indication of track loss). The trials were also rejected when the first fixation fell on the defined AOI earlier than 100 ms after picture onset because eye movements in response to stimulus presentation are unlikely to occur this early (Duchowski, 2007). These exclusions affected 15% of the data in Experiment I and 11% of the data in Experiment II (see **Table 1** for the number of exclusions by condition).

| Distractor | Sep. attr. | Experiment I | Experiment II | Total |
|---|---|---|---|---|
| absent | absent | 91 | 106 | 197 |
| absent | present | 103 | 102 | 205 |
| present | absent | 119 | 123 | 242 |
| present | present | 105 | 95 | 200 |
| Total | | 418 | 426 | 844 |

**Table 1:** Number of excluded trials in Experiment I and Experiment II as a function of the four conditions.

All fluent descriptions of pictures were scored for their syntactic structure as transitive sentences (e.g., *The man is pushing the car*), passive sentences (e.g., *The car is pushed by the man*), truncated passives (e.g., *The car is pushed*), intransitive sentences (verbs followed by nouns carrying oblique case marking in Estonian; e.g., *Laps mäng-ib karu-ga*, child.SG.NOM play-3SG.PRS bear-SG.COM, 'The/a child is playing with the/a bear') or other constructions. Only transitive sentences were included for further analysis. Within the subset of transitive sentences, further utterances with speech onset latencies of more than three standard deviations above the grand mean were excluded.

Furthermore, the mentions of inseparable features were scored as short modifiers and the mentions of separable features as long modifiers as they were expected to differ in length (adjective phrases vs. modifier noun phrases, see Examples in 1). In Experiment I, speakers produced plain nouns or the so-called zero modifiers, short modifiers and long modifiers as expected, but they also mentioned some other modifiers and generated combinations of different modifier types (see **Table 2**). Speakers produced considerably fewer modifiers in Experiment II than in Experiment I. In particular, in Experiment I, the unexpected modifiers and modifier combinations constituted 15% of all modifiers mentioned in transitive sentences, whereas in Experiment II, they constituted only 1.7% of all mentioned modifiers.

| | | Experiment I | | | | | Experiment II | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Modifier** | **Distractor:** | **absent** | **absent** | **present** | **present** | **Total** | **absent** | **absent** | **present** | **present** | **Total** |
| | **Sep. attr.:** | **absent** | **present** | **absent** | **present** | | **absent** | **present** | **absent** | **present** | |
| zero | | **468** | **422** | 32 | 54 | 976 | **812** | **791** | **766** | **742** | 3111 |
| short | | 58 | 26 | **442** | 8 | 534 | 3 | 1 | 5 | 2 | 11 |
| long | | 10 | 86 | 67 | **476** | 639 | 0 | 9 | 0 | 16 | 25 |
| other | | 17 | 17 | 12 | 2 | 48 | 3 | 5 | 6 | 3 | 17 |
| short + short | | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| short + long | | 0 | 1 | 0 | 6 | 7 | 0 | 0 | 0 | 0 | 0 |
| short + other | | 0 | 0 | 3 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| long + other | | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| **Total** | | 553 | 553 | 557 | 547 | 2210 | 818 | 806 | 777 | 763 | 3164 |

**Table 2:** Frequencies of different modifier types in transitive sentences as a function of the four experimental conditions in Experiment I and Experiment II (*Zero* modifiers refer to mentions of patients without any attributes, *short* modifiers refer to mentions of inseparable attributes, *long* modifiers refer to mentions of separable attributes and *other* modifiers refer to mentions of unexpected modifiers). The numbers in bold indicate the data that was evaluated for the results.

Overall, the data in **Table 2** demonstrate that the two different task settings in the experiments triggered expected specifications of the two interacting actors. For the analysis of the results, utterances that contained plain nouns or the so-called *zero modifiers*, adjective phrases or the so-called *short modifiers*, and modifier noun phrases or the so-called *long modifiers* were evaluated in their respective conditions (see the numbers in bold in **Table 2**). Transitive sentences with unexpected modifiers and modifier combinations were not included in the analysis.

Finally, all utterances longer than three standard deviations above the grand mean sentence duration were excluded. Thus, the final set from Experiment I consisted of 54% of the sentence productions (1557 tokens), and the one from Experiment II consisted of 65% of the sentence productions (2648 tokens). The amount of exclusions is characteristic for this type of spontaneous speech production task where speakers are free to produce any kind of picture descriptions. In addition, Estonian is rich in non-transitive sentence constructions that enabled the speakers to mention both interacting actors/objects in the pictures. Although they were sufficient as informative picture descriptions, the non-transitive sentences were excluded based on the transitivity criterion. The exclusions, specifically those based on the transitivity of the sentences (excluding constructions other than transitive sentences) and the timing and duration properties of the utterances (i.e., speech onsets and duration of sentences), are crucial for the relative comparability of the time course of planning in these utterances.

## 2.5 Analysis

Acoustic analyses were carried out with the help of a digital tool for phonetic analyses (Praat, Boersma and Weenink 2020). Every utterance was manually tagged for the starting time of the first word and the ending time of the last word. Based on the transcripts, the timestamps of words, hesitations and prosodic breaks were automatically identified with the help of a web tool for automatic alignment (WebMAUS, Kisler, Reichel, and Schiel 2017). The forced alignment of the word boundaries was manually corrected, and the words were additionally annotated for the onsets and offsets of the agent names and patient names. Based on this data, the speech onset latencies and the durations of spoken sentences were automatically extracted from the speech signals.

F0 contours of utterances were created using the autocorrelation method available in Praat in two passes. During the first pass, F0 contours were extracted with default settings for the lowest and highest F0, the so-called floor and ceiling (75 Hz and 600 Hz, respectively). Then, for each speaker, the average first quartile and third quartile of F0 (Q1 and Q3) was calculated across all the speakers' F0 contours. In the second pass, F0 contours were created with speaker-specific settings (0.75*Q1 for floor and 1.5*Q3 for ceiling). F0 maxima from the first words of the sentence-initial nominal constituents, i.e., the subject noun phrases encoding the agent, were automatically measured from the extracted F0 contours (see **Figure 2**) and submitted for further analysis.

(3) a. pruut lõika-b tort-i
bride.*SG.NOM* cut-*3SG* cake.*SG-PART*
'a/the bride is cutting a/the cake'

b. naine süüta-b päkapiku-ga küünla
woman.*SG.NOM* light-*3SG* elf.*SG-COM* candle.*SG.PART*
'a/the woman is lighting a/the candle with [a/the picture of] an/the elf'

To normalize for the gender differences in voice pitch, F0 maxima were converted into semitones based on the speaker mean F0 ($F0_{ref}$) averaged across all utterances of a speaker (see Formula 1).

$$F0_{st} = 12 * log_2(\frac{F0}{F0_{ref}}) \tag{1}$$

The distribution of F0 maxima (in semitones) demonstrated a number of outliers (very long thin tails in the density curve). For the regression analysis, the pitch measurements were additionally trimmed by detecting the residual outliers. In particular, the linear mixed effects regression analysis was run with a full random effects structure (see **Table 5** in Appendix 2). Subsequently, any pitch measurements that demonstrated residuals more than two standard deviations away from the mean residual were excluded from the sample (and the model was run again).

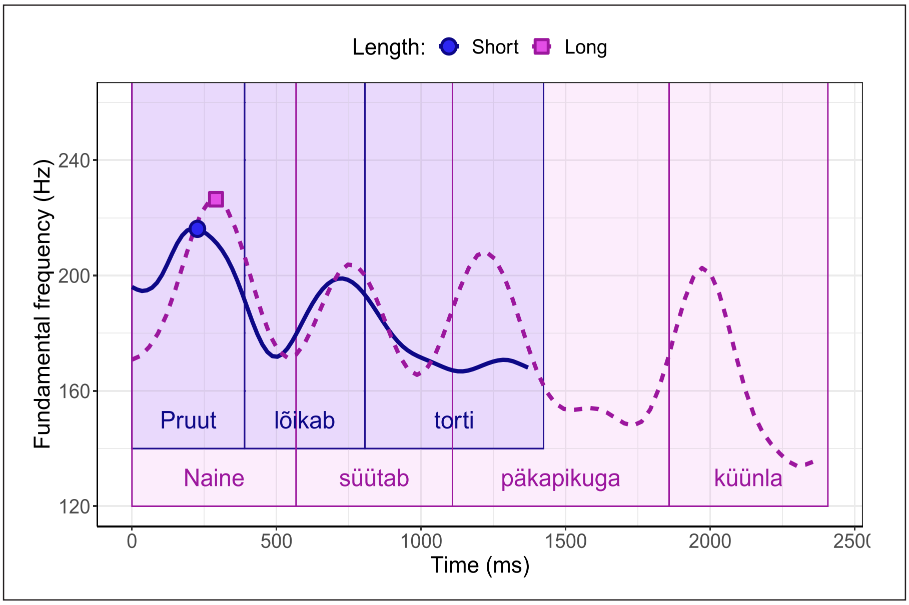

**Figure 2:** Fundamental frequency contours (in Herz) of short and long utterances from a female speaker (solid and dashed lines, respectively; see 3a and 3b for translations and glosses). Vertical lines refer to word boundaries. The points on the contour highlight the highest F0 (F0 maximum) extracted from the very first words of short and long utterances (circle for short, rectangle for long).

For Experiment I, the picture descriptions were also annotated for prosodic chunk boundaries. The chunk boundary was defined by the occurrence of a silent or filled pause, hesitation and partly by pre-boundary lengthening. The picture descriptions were examined for major disfluencies, and multi-chunk utterances were included if they still conveyed the expected content fluently. Pauses and hesitations are common in spontaneous speech data. Arguably, the inclusion of multi-chunk utterances provides a stronger generalization of the length effect on sentence intonation. Specifically, F0 maxima are expected to occur more frequently in longer picture descriptions. When a longer picture description is divided into two or three shorter chunks, the length of the first chunk is inevitably shorter than the expected short utterance. According to the length effect, the intonation peaks should then occur lower in longer utterances than in shorter ones. However, if the F0 measurements counteract this expectation, the results would strongly suggest that the length-dependent scaling of intonation peaks operates across the utterance content and that the length effect generalizes over the variation of prosodic chunk boundaries.

**Figure 3** indicates that the number of chunks in the fluent utterances increased as the linguistic complexity increased. In particular, the proportion of single-chunk utterances decreased from 66% to 38% of all the utterances through the four conditions in Experiment I. A preliminary exploration of the data did not suggest that the utterance-initial F0 maxima were dependent on the number of prosodic chunks. Therefore, the results of Experiment II did not undergo such careful prosodic analysis, and the information about the number of prosodic chunks or the duration of the utterance-initial prosodic chunks was not included in the statistical analysis.

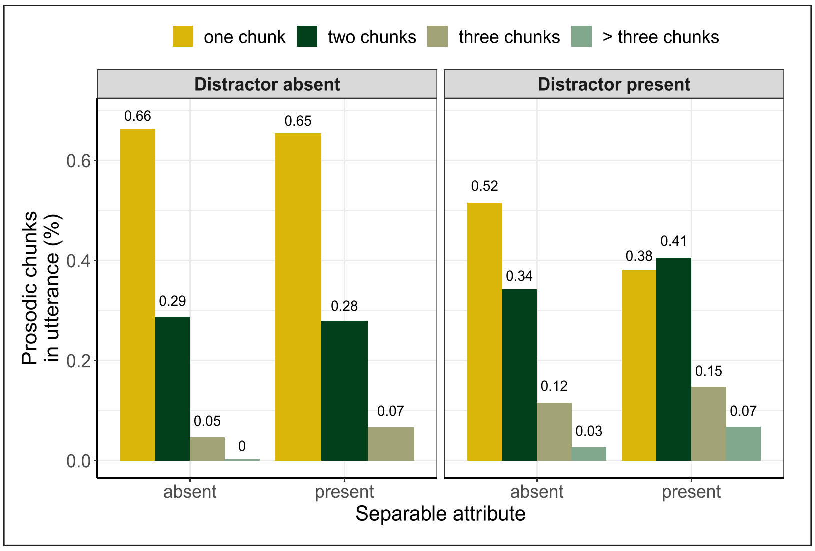

**Figure 3:** Proportions of utterances constituting a single prosodic chunk (yellow), two prosodic chunks (dark green), three prosodic chunks (light green) and more than three prosodic chunks (blue) as a function of factors Distractor (absent or present) and Separable Attribute (absent or present) in Experiment I.

For the evaluation of eye movements, multiple consecutive fixations on corresponding AOIs were combined into gazes, each with individual onset, offset and duration (see also Griffin & Davison, 2011). The analysis concentrates on a naming gaze, which was defined as the longest gaze occurring before the start of speech. In contrast, studies of eye movements assess the averaged proportions of gaze as a function of time (Bock et al., 2004; Konopka & Meyer, 2014; Sauppe, 2017) (see also **Figure 7d** and **8d** in Appendix 1). These analyses of sentence production usually determine the fixed time frame for the conceptualization stage (usually 400 or 600 ms after picture presentation) and determine the effects of linguistic factors on gazes with the help of growth curve analysis (Mirman, 2014). Alternatively, the analysis may assess the duration of some critical gazes in relation to speech onset latency (Kuchinsky et al., 2011; Swets, Fuchs, Krivokapić, & Petrone, 2021). For instance, Swets et al. (2021) assessed the scope of message generation with the proportional measure of gaze. In their visual world picture naming study, the critical AOIs were expected to be mentioned only later in picture descriptions. Early eye gazes to these critical areas were taken to index the large scope of message planning such that the increasing duration proportion of gazing at this critical area indexed the scope of utterance planning (how far in advance the utterance was planned).

In my approach, instead of selecting a critical AOI, I utilize the duration proportion of a gaze with the longest duration (i.e., the relative duration of the naming gaze). To achieve this, gazes of maximum duration were identified within the time frame from picture onset to speech onset and served as the so-called naming gazes (see **Figure 4**). Based on earlier evidence on sentence planning (Bock et al., 2004; Griffin, 2001; Griffin & Bock, 2000; Levelt et al., 1991), every trial is expected to exhibit a gaze of longest duration, indicating lemma retrieval for first-mentioned actors/objects in the corresponding pictures. The start and duration of these longest naming gazes signal that the brief apprehension of event gist has passed. The duration proportion, relative to speech onset latency, captures these brief apprehension times. As the focus of pre-speech visual processing shifts away from the agents, the apprehension of the event gist involves actors other than the agent, indicating a large scope of message generation. Thus, larger increments of conceptual planning are reflected in shorter relative durations of naming gazes, while longer relative durations indicate smaller increments of conceptual planning.

## 2.6 Evaluation

For the investigation of conceptual planning, a linear mixed effects regression (LMER) analysis modelled the relative duration of the naming gaze as a function of a three-way interaction between Distractor (absent vs. present), Separable Attribute (absent vs. present) and Experiment.

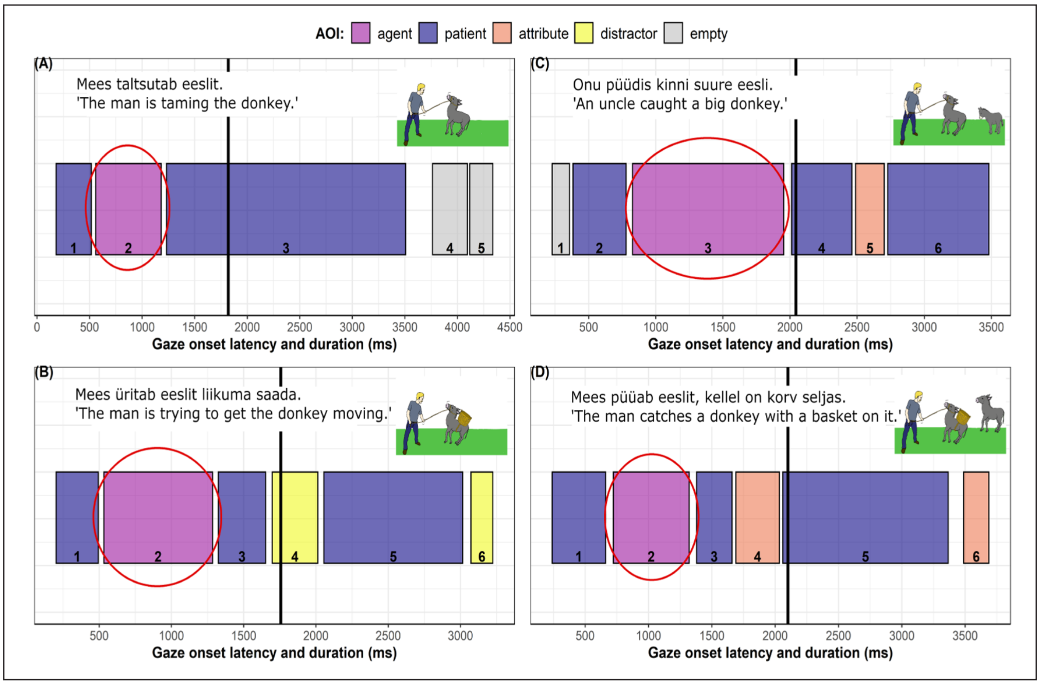

**Figure 4:** Gaze onset latencies and durations of four Estonian speakers describing the item "man pulling" in four different conditions of Experiment I, which was designed to elicit descriptions varying in length and conceptual complexity. The positions and the widths of the squares indicate when and for how long a gaze was directed towards a particular area of interest (e.g., purple refers to the agent initiating and blue to the patient undergoing the action of pulling; pink and yellow refer to the attribute and the distractor, respectively; the grey boxes refer to gazes directed outside of the defined AOIs). The numbers on the squares count the gazes. The black vertical lines indicate speech onset latencies. The red ellipses highlight the gazes that are called "naming gazes" and are defined as gazes of the longest duration directed towards the AOIs of first-mentioned actors before the starts of the utterances.

The random effects allowed the slope and intercept of the three-way interaction to vary by items and the slope and intercept of the two-way interaction to vary by participants as long as the model converged. To tackle convergence issues, the correlation between the items and subjects was first removed. When this did not help, the random effects structure was simplified by excluding the interaction between the predictors. If these measures did not help, the slopes were removed one by one.

The significance of the fixed effects of the converging model was obtained by performing likelihood ratio tests with the help of the *afex* package (Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2022). In particular, the *p*-values reported below are the results of the likelihood ratio tests of two nested models where one of the models contained the predictor of interest and the other model excluded the predictor (for instructions for testing hypotheses with the help of linear mixed effects regression methods, see, e.g., Speekenbrink, 2022; Winter, 2019). The *p*-values of the likelihood ratio tests express the credibility of the fixed effect, that is, the confidence that the parameter would be observed again in the future given the data.

For testing the relationship between the advance planning of sentence intonation and the incrementality of conceptual planning, the relative duration of the naming gaze was redefined as a continuous predictor variable of utterance-initial F0 maxima. For the phonological component, the analysis included Experiment as a categorical predictor variable. Thus, the LMER analysis modelled the height of sentence-initial intonation peaks as a function of a two-way interaction between the relative duration of the naming gaze and Experiment. The continuous variables F0 maxima and the relative durations of the naming gazes were centered and scaled for the regression fitting. As the model evaluated the effect of one continuous predictor and a between-subject categorical factor, the random effects structure included only random intercepts for items, participants and trial number. Similar to the analysis above, the credibility of the parameters is reported based on the likelihood ratio tests obtained with the help of the *afex* package (Singmann et al., 2022).

## 3. Results

The results were evaluated from the two related perspectives of planning sentences and sentence intonation in Estonian. Firstly, the aim was to investigate how incrementally or non-incrementally Estonian sentence planning may proceed (the flexibility of conceptual planning). Secondly, the goal of the evaluation was to detect whether and how the advance planning of sentence intonation relates to the conceptual and/or linguistic planning processes.

### 3.1 Conceptual and linguistic factors of sentence planning in Estonian

The investigation of the relative gaze duration in this subsection is expected to shed light on the sensitivity of conceptual planning to the categorical factors of conceptual and linguistic complexity. Recall that both factors Distractor and Separable Attribute modulated the conceptual complexity of the picture descriptions. Similarly, both factors were responsible for longer rather than shorter picture descriptions in Experiment I. Nevertheless, the factor Distractor is regarded as a linguistic factor because the presence of a distractor cued the necessity for linguistically complex or semantically more concrete references to the patients in the pictures of events. The factor Separable Attribute is regarded as a conceptual factor because inclusion of separable attributes did not always trigger longer rather than shorter sentences (see **Figures 4A** and **4B**). The hypothesis predicts additive effects of conceptual factor (i.e., Separable Attribute) and linguistic factor (i.e., Distractor) on the relative duration of the naming gaze irrespective of Experiment.

The results in **Figure 5** show that the relative gaze duration is greater for the conditions where the distractors were absent than for the conditions where the distractors were present. This visual observation shows the effect of Distractor on the relative duration of the naming gaze. Similarly, the relative duration of the naming gaze appears greater for the conditions where the inseparable attributes were absent in the pictures than for the conditions where the inseparable attributes were present. This suggests an effect of Separable Attribute on the relative duration of the naming gaze. The declining trend of the means of the relative gaze durations across the four conditions appears similar for Experiment I and Experiment II.

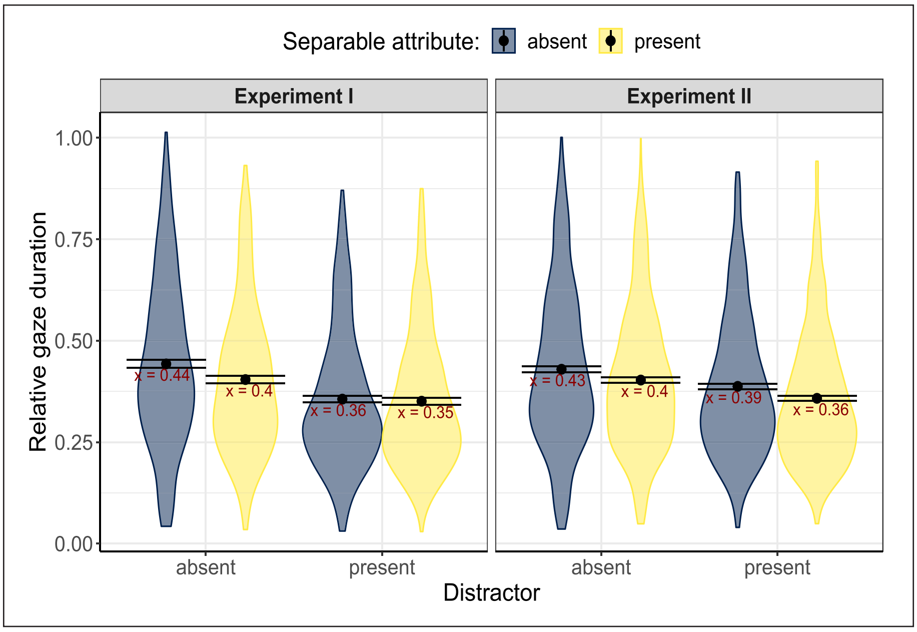

**Figure 5:** The distributions of the relative duration of the naming gaze across the two experiments as a function of the factors Distractor (absent vs. present) and Separable Attribute (absent vs. present). The condition in which a separable attribute and a distractor are absent (the grey colour indicates the absence and the yellow colour the presence of separable attributes) is the simplest, containing just the two interacting actors, whereas the condition with a separable attribute and a distractor present is the most complex, containing three actors and an object as a characteristic feature of the interacting patient (see **Figure 1**). The black labelled circles indicate the means of the variables and the error bars represent the 95% confidence intervals. The LMER analysis reports significant differences as long as the 95% confidence intervals do not overlap.

The LMER analysis confirms that the relative duration of the naming gaze becomes shorter as the distractors and inseparable attributes are included in the pictures, i.e., the size of the visual array increases. The likelihood ratio tests indicated a significant main effect of Distractor ($\chi^2(1) = 96.41, p < 0.001$) and Separable Attribute ($\chi^2(1) = 14.43, p < 0.001$). No significant main effect of Experiment occurred. There were no significant three-way and two-way interactions between the factors Distractor, Separable Attribute and Experiment (for the LMER summary, see **Table 4** in Appendix 2).

The main effects of Distractor and Separable Attribute indicate that the relative duration of the naming gaze decreases in the presence of a distractor and/or a separable attribute. The model parameters in **Table 4** (Appendix 2) demonstrate that the decrease in the relative duration of the naming gaze is larger for Distractor than for Separable Attribute, which can also be observed in the gradually decreasing relative duration of the naming gaze across the four conditions in **Figure 5**. The lack of a significant interaction between Distractor, Separable Attribute and Experiment indicates that the decreasing trend does not differ for the two experiments. Thus, the data provides support for the additive effects of conceptual and linguistic complexity on message generation irrespective of the two experiments.

Regarding the incrementality of conceptualization, the relative gaze durations indicate an unexpectedly large scope of message planning. Namely, the greatest proportion of gazing agents (i.e., 44%) occurs for the simplest condition (attribute absent, distractor absent) in Experiment I. All proportions remain below 50% and indicate that for more than the half of initial processing time, speakers allocate their visual attention to event participants other than agents. These results suggest only weakly incremental conceptual planning for Estonian. The statistical significance of the 9% change in Experiment I and 7% change in Experiment II nevertheless suggests that there was still space to vary the scope of message generation.

## 3.2 The relationship between sentence planning and intonation planning in Estonian

This subsection investigates whether the advance planning of sentence intonation in Estonian relates to different degrees of incrementality in conceptual planning or to the length of utterances. Recall that the manipulation of the linguistic and conceptual complexity (factors Distractor and Separable Attribute) was expected to modulate the scope of conceptual planning such that the planning increments for simpler conditions would contain only concepts of agents and the planning increments for more complex conditions would include concepts of patients as well. Consequently, the different degrees of incrementality of conceptual planning were approximated by the relative duration of the naming gaze. The length of utterances was manipulated across the two experiments. While speakers produced short and long picture descriptions in Experiment I, they produced only short picture descriptions in Experiment II. For the intonation planning to be associated with the conceptual planning, the hypothesis predicts a positive correlation between the relative duration of the naming gaze and the sentence-initial intonation peaks, irrespective of Experiment. Alternatively, the sentence-initial intonation peaks might relate to the length of the sentences only (a main effect of Experiment) or to both the relative duration of the naming gaze and the sentence length.

**Figure 6** reveals no effect of the relative duration of the naming gaze on the sentence-initial intonation peaks. Nevertheless, the effect of length on the sentence-initial intonation peaks can be observed such that they appear somewhat higher in Experiment I than in Experiment II. In addition, **Figure 6** suggests a tendency for F0 maxima to increase as the utterance duration increases (an additional linear mixed effects model with the interaction between duration and Experiment reported a significant main effect of duration [$\chi^2(1) = 51.39, p < 0.001$]; see **Table 3** in Appendix 2).

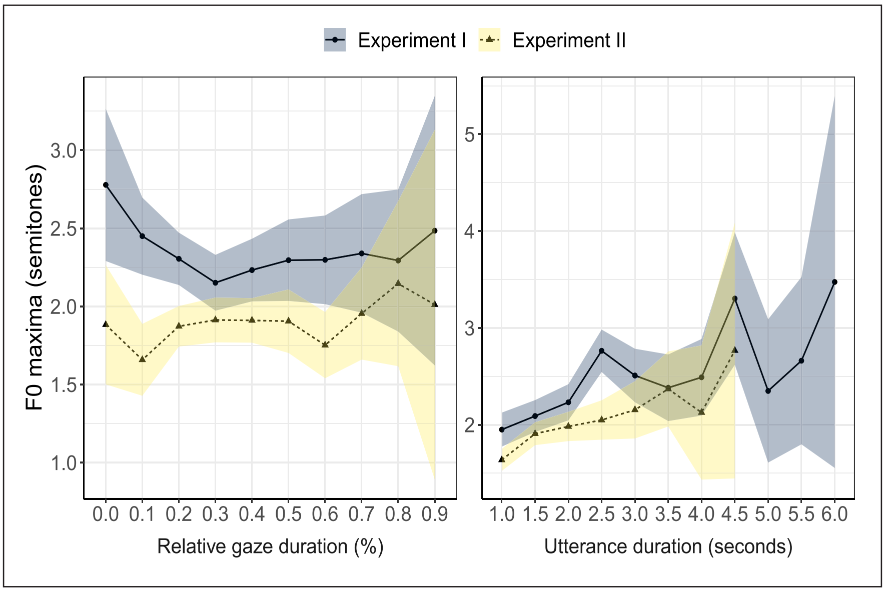

**Figure 6:** The average height of sentence-initial intonation peaks (in semitones) as a function of the relative duration of the naming gaze (the gaze duration as percentage of the speech onset latency) and the duration of utterances (seconds) for Experiment I (indicated by circles) and Experiment II (indicated by triangles). F0 maxima were averaged by 0.5-second bins of the relative gaze duration and utterance duration. The shaded areas around the means of F0 represent the 95% confidence intervals (grey indicates Experiment I and yellow refers to Experiment II).

The LMER analysis of the F0 maxima as a function of the interaction between the relative gaze duration and Experiment confirms the effect of length on the sentence-initial intonation peaks. The likelihood ratio tests reported a significant main effect of Experiment ($\chi^2(1) = 6.06$, $p < 0.05$; estimated increase = 0.123 semitones) on the sentence-initial F0 maxima. There was no effect of the relative duration of the naming gaze and no significant interaction between the relative duration of the naming gaze and Experiment (for the LMER summary, see **Table 5** in Appendix 2).

The main effect of Experiment indicates that the production of sentence intonation was highly sensitive to the length of the picture descriptions (in milliseconds). This means that the length of utterances contributes to the height of utterance-initial intonation peaks such that the intonation peaks increase together with the increasing length of the utterances (consider the shorter and longer picture descriptions in Experiment I against the short picture descriptions in Experiment II). Surprisingly, the height of the utterance-initial intonation peaks did not depend on the relative duration of the naming gaze, that is, the size of the increment of conceptual planning. Thus, the results provide no support for the hypothesis that the tonal scaling associates to the incrementality of conceptual planning. The significant effect of Experiment demonstrates the role of linguistic representation of incipient utterances in intonation planning.

## 4. Discussion

The results of the two visual world picture description experiments indicate that the conceptual planning processes in Estonian are somewhat affected by the different degrees of conceptual and linguistic complexity. In particular, with increasing visual complexity, speakers tended to distribute their attention between the actors/objects more evenly and to gaze at the agents (the first-mentioned actors/objects) for shorter time proportions of the speech onset latencies. A closer look into the eye gazes, nevertheless, suggests a large increment of conceptual planning irrespective of complexity manipulations. Namely, the amount of gazing at the agents in the simplest condition (distractor absent, inseparable attribute absent) constituted just about 44% of the speech onset latency. This means that half of the conceptual processing time was spent for the processing of the patients already in the simplest speaking conditions which, in turn, suggests an overall preference for the less incremental, also referred to as relational generation of messages. Although the changes of 9% in Experiment I and 7% in Experiment II are small, their statistical significance indicates that the scope of message planning was sensitive to the conceptual and linguistic complexity of the subsequent utterances.

In addition, the results revealed that the tonal scaling of utterance-initial intonation peaks differed between the two experiments. Intonation peaks occurred higher in Experiment I compared to Experiment II. To recap, Experiment I involved an increase in the length of picture descriptions across the four conditions, while Experiment II maintained a constant utterance length. Consequently, the tonal scaling of sentence-initial intonation peaks was influenced by the linguistic complexity of the utterances. Therefore, the findings of the current study suggest a strong presence of intonation planning in Estonian. However, the experiment was not able to detect an effect of conceptual planning scope on the height of utterance-initial intonation peaks. The 9% or 7% change in relative gaze duration may have been too subtle to significantly impact the height of these intonation peaks. Nevertheless, the overall preference for a larger planning scope in messages and the persistent length effect on intonation peaks collectively indicate that successful intonation planning occurs under the extended scope of message generation.

## 4.1 Sentence planning in Estonian

Estonian is a morphological case-marking language (Viitso, 2003) that exhibits a relatively free ordering of sentence constituents (see, e.g., Erelt, 2003; Erelt & Metslang, 2017; Vilkuna, 1998). These parameters create an expectation for a less incremental conceptual planning stage (Norcliffe et al., 2015; Sauppe, 2017). Indeed, 44% or less speech onset latency was allocated to the agent. In other words, the patients and the other actors in the pictures were visually processed for 56% of the speech onset latency or even longer. Thus, the relative duration of the naming gaze indicated a rather less incremental conceptual planning across all four conditions in both experiments. These results suggest that Estonian speakers prefer to generate a more comprehensive rather than a limited relational framework of subsequent utterances before they start the grammatical encoding of their messages. A more comprehensive conceptual framework might be necessary because the speech production system in Estonian may be anticipating the encoding of grammatical case. Namely, the patients that were mentioned after agents in the experiments of this study were marked for the partitive case (see Example 1 in Section 2), which in Estonian frequently encodes semantic event patients or syntactic objects. The results suggest that Estonian speakers might target the event patients early at the outset of picture processing because their grammatical function needs to be assigned together with particular case suffixes. According to Bock and Levelt (1994), the closed class elements such as inflectional (and declinational) suffixes are accessed separately from stem forms (lemmas). Thus, at the time of conceptual planning, the Estonian production system may anticipate the retrieval of case suffixes, in addition to lemmas.

The speech onset latencies, though not reported in the Results, were also quite long for Estonian speakers (see **Figure 9** in Appendix 3). In earlier picture description studies, the speech onset latencies for speakers of Germanic languages ranged between 1800 and 2000 ms after picture presentation (see, e.g., Konopka & Meyer, 2014; Sauppe, 2017; van de Velde et al., 2014). The average speech onset latency for Estonian speakers in simple as well as in more complex picture descriptions was longer than 2100 ms after onset of the visual display. To ensure a better cross-linguistic comparison, the participants of Experiment II were even asked to start speaking faster when they appeared to be slow in performing the task. Despite this, the speech onset latencies of Experiment II mirror the speech onset latencies of Experiment I well. Relatedly, a study of another grammatically complex language has found strikingly long speech onset latencies as well. Namely, Myachykov, Scheepers, Garrod, Thompson, and Fedorova (2013) found that Russian speakers start speaking extremely slowly when compared to English speakers. Myachykov et al. (2013) argue that languages with relatively flexible word orders encompass a greater choice of sentence types for encoding messages compared to languages with restricted word order, and they suggest that the greater structural diversity might slow down the planning and production processes. This might be the case for Estonian as well. Similar to Russian, Estonian exhibits free ordering of sentence constituents and a variety of sentence types in addition to the transitive structure. Therefore, the sentence planning processes at the outset of grammatical encoding might be occupied with structural competition. Moreover, the structural diversity might not only slow down the encoding and production processes, but also require relational encoding at the stage of conceptualization for all these alternative types of sentence structures to become activated and be considered.

Thus, both the relative duration of the naming gaze and the speech onset latencies of the two experiments indicate that Estonian speakers plan more than just the concept of agent, that is, the utterance-initial constituent. They might plan the concept of the event in terms of a very abstract category; in the current case, a category of *action* with a function of CAUSE (see Levelt, 1989, 78–81). The recognition of the function of CAUSE also activates on an abstract level a thematic structure that requires two semantic arguments. Thus, at the time of conceptualizing, the speakers might already be anticipating encoding the two semantic arguments. In addition, speakers might conceptualize some abstract properties of these arguments, such as something about their state in the action or animacy. They may already identify the cause and the target of the action (i.e., who is doing what to whom). The rapid recovery of this kind of relational structure might be key to ensuring that the relevant case suffixes are retrieved fast enough during the formulation processes.

Notably, in Estonian, the case suffixes attach to the stem forms more irregularly than the endings *-s* or *-ed* in English. Often, stem alternations are necessary to attach the case suffix correctly (see, e.g., Viitso, 2003). For example, the declension of the word ‘door’ in Estonian involves two different word stems: *uks* for nominative case and *us-t* (door.SG-PART) for partitive case (sentential objects are often marked for partitives in Estonian). Therefore, Estonian speakers might be involved in lexical encoding (see Levelt, 1989) rather than retrieval of lemmas and suffixes, but this topic is clearly beyond the scope of the current study. More importantly, the results support the idea that speakers of a case-morphological language conceptualize messages less or non-incrementally. The non-incremental conceptualization processes and the resulting comprehensive conceptual framework might reflect the difference between the message encoding in case-marking and non-case-marking languages. Moreover, it might be responsible for the missing effects of lexical accessibility in a number of other morphological case-marking languages, such as Finnish, Korean and Russian (Hwang & Kaiser, 2015; Myachykov & Garrod, 2008; Myachykov & Tomlin, 2008). In addition, the results of the study appear to suggest that while the relational planning of messages is a choice of the speech planning strategy in non-case-marking languages (see, e.g., Konopka & Meyer, 2014), the relational conceptual planning is required in the case-marking languages.

Regarding the varying size of the planning scope, the statistically significant effects of linguistic and conceptual complexity on the relative duration of the naming gazes indicate that there is still space for the conceptual planning scope to vary in situations and languages where, overall, a less incremental planning strategy is involved. Specifically, as conceptual and linguistic complexity increased, the proportion of gazing at agents decreased by approximately 9% and 7% of the speech onset latency through the four experimental conditions in the first and second experiments, respectively. Most likely, in these more complex experimental conditions, the scope of conceptual planning involved an additional dimension of referential situations. Namely, the conceptually and linguistically complex speaking situations required speakers to assess the contrast between the two similar actors/objects (Experiment I) or between the members of the same group (Experiment II). For efficient linguistic encoding, the conceptual planning processes might have needed to identify and store information about the contrast. Thus, the abstract conceptual framework acquired an additional dimension, which was stored and made accessible for the subsequent procedures of grammatical encoding.

## 4.2 The advance planning of sentence intonation

With regard to intonation planning, the two experiments differed in terms of language planning and production processes. In Experiment I, the linguistic complexity of the utterances (short vs. long) varied in conjunction with conceptual complexity, while linguistic complexity was controlled for in Experiment II. Specifically, in Experiment II, speakers produced short picture descriptions but the degree of conceptual complexity varied similar to Experiment I. The comparison of the height of sentence-initial intonation peaks across the two experiments showed that the tonal scaling of sentence-initial intonation peaks was sensitive to the linguistic complexity and not to the conceptual complexity. In other words, utterance-initial intonation peaks increase together with increasing length of utterances, but not with increasing scope of conceptual planning. The length-dependent scaling of sentence-initial intonation peaks aligns with several established findings (Cooper & Sorensen, 1981; Liberman & Pierrehumbert, 1984; Prieto et al., 2006; Thorsen, 1985; Yuan & Liberman, 2014) and supports the concept of advance planning of sentence intonation.

This result also corroborates the length effect on intonation peaks in both read-aloud and in spontaneous Estonian (see Asu, Lippus, Sahkai, and Salveste (2017) and Asu et al. (2016), respectively). In Asu et al. (2016), the length effect occurred in the prosodic chunks that were defined by the presence of a prosodic break (i.e., hesitation, pause, segmental lengthening). In their study, the relationship between the utterance content and prosodic chunking was left open. In the current study, an utterance is defined by its content. The initial boundary is determined with reference to agent of the action, and the final boundary necessarily includes the reference to the patient. Despite being relatively fluent, these content-driven utterances were separated into several prosodic chunks (see **Figure 3** in Section 2.5). The fact that the higher intonation peaks still occurred in longer utterances indicates that the length-dependent scaling of utterance-initial intonation peaks generalizes over multi-chunk utterances, suggesting a strong presence of intonation planning in Estonian. Further studies are necessary to scrutinize the cross-linguistic generalization of this finding.

When replicated, the finding on length-dependent scaling of sentence-initial intonation peaks clearly constitutes compelling evidence against incremental planning of sentence prosody (Levelt, 1989; Levelt et al., 1999), possibly not only for Estonian but also for other languages. In his sketch for connected speech, Levelt (1989) argues that intonational parameters can be set with a minimum lookahead of one word, and a prosodic break is proposed to be the main vehicle to enable incrementality in intonation planning. The model proposes that the declination trend, of which the parameter of sentence-initial intonation peak is a part, is rather a consequence of running out of breath or dropping the voice pitch too rapidly. In these difficult cases of intonating, a speaker can opt for the break option at any time and at the end of the last articulated word at the latest. Thus, the production of prosodic breaks in Levelt (1989) is entirely an incremental strategy, and this has been implemented to argue against a number of prosodic phenomena that have been suggested to require some lookahead during speech production (e.g., stress clash in English and final lengthening). With respect to this notion of incremental planning of sentence prosody in connected speech, the current study raises a highly intriguing question. Namely, if a prosodic break enables an option to delay sentence planning or to avoid an inconveniently low level of voice pitch, why is the parameter for sentence-initial intonation peaks still set with reference to the length of incipient utterances, that is, with a lookahead of a whole utterance?

One tentative explanation is that speakers might aim to deliver a message as a single tonally coherent unit (for the notion of tonal coherence, see Bois, Cumming, Schuetze-Coburn, & Paolino, 1992; Breen, Fedorenko, Wagner, & Gibson, 2010; Buhmann et al., 2002; Himmelmann, Sandler, Strunk, & Unterladstetter, 2018). The prosodic phenomena that underlie the percept of tonal coherence are the speech rhythm, a regular placement of pre-nuclear and nuclear pitch accents and the continuously declining or rising trend of F0. Perceptually, a prosodic break together with a pitch reset would disrupt the intended tonal coherence of a message. Therefore, Estonian speakers of the current study might have aimed to fit the entire content of their utterances within a single continuous F0 trend that spanned several prosodic chunks, which, in turn, constituted a transitive sentence. This means that the declination trend that began at the sentence-initial agent name continued after a prosodic break in the subsequent prosodic chunk through the end of the utterance. To achieve this, they set a higher parameter for the sentence-initial intonation peaks as soon as they anticipated the production of modified noun phrases. This type of scaling of intonation peaks requires a larger scope of lookahead than one word.

From a cross-linguistic perspective, the regularity of length-dependent scaling of sentence-initial intonation peaks in Estonian is somewhat surprising. Although the length-dependent scaling of intonation peaks was already noted long ago and occurs in a number of languages (for English, see Cooper and Sorensen 1981; Liberman and Pierrehumbert 1984; Yuan and Liberman 2014; for Danish, see Thorsen 1980; for Mandarin Chinese, see Yuan and Liberman 2014), it appears less regularly or is even absent in a number of other languages (for German, see Fuchs et al. 2013; for a comparative study of Romance languages, see Prieto et al. 2006; for Danish, see Tøndering 2011). For example, Fuchs et al. (2013) reports for read-aloud German that the height of sentence-initial intonation peaks was sensitive to the length of sentence-initial constituents, but it was insensitive to the length of sentence-final constituents. Based on this result, Fuchs et al. (2013) argue that the advance planning of sentence intonation is limited to sentence-initial constituents in German. The problem with most of these studies on intonation planning is that they examine read-aloud sentences with highly controlled phonological structures. Reading aloud is necessarily a linear process with a varying size of lookahead. A preview of upcoming linguistic material whilst reading might include the whole sentence when the first constituent is relatively easy. However, the preview might include only the first constituent when it is phonologically more complex. Thus, the longer and more complex actor names at the beginning of written sentences in Fuchs et al. (2013) might have triggered the smaller units of phonological planning. Notably, a couple of planning components for spontaneously produced utterances proceed non-linearly. For instance, the emergence of a message is non-linear or several lemmas can become activated simultaneously. Thus, especially for prosodic planning, the generalization from the read speech to the spontaneous speech may be particularly challenging. More surprising and interesting is the fact that in Estonian, the length-dependent scaling of sentence-initial intonation peaks holds for both read and spontaneous speech. In this respect, further studies comparing different speaking modes and languages would certainly deliver interesting patterns of results and enable more conclusive generalizations about the advance planning of sentence intonation.

### 4.3 Intonation planning and sentence planning in Estonian

Subsection 4.1 concluded that Estonian speakers prefer to plan their messages in larger rather than smaller planning increments. Within the larger planning scope, there was still room for an increase in the scope of conceptual planning. This increase in planning scope was attributed to integration of the contrast between the two potential referents in the pre-verbal message. However, the increase in planning scope turned out to be too small to exercise an effect on the utterance-initial intonation peaks. Instead, the sentence-initial intonation peaks were sensitive to the variations in utterance length across the two experiments. As the two experiments differed in terms of linguistic complexity, the results suggest that it was the linguistic complexity of the utterances that mattered the most for the tonal scaling at the beginning of the spontaneous utterances. As mentioned earlier, the linguistic factor was applied to probe for the linguistic planning processes (see the introduction of Section 2). Therefore, the results warrant the conclusion that intonation planning relies on the linguistic encoding processes. Consequently, the question arises of what type of information at the stage of linguistic encoding, in particular the grammatical and phonological encoding, underlies the intonation planning in Estonian.

The experiments of the current study indicated that speakers of Estonian successfully anticipated the linguistic complexity of the patients, i.e. the weight of the last-mentioned noun phrases (bare nouns vs. modified noun phrases), and planned the height of sentence-initial intonation peaks accordingly. On the one hand, the weight of the noun phrase means a longer string of words (an adjective plus a noun in the current case). The intonation planning could rely on this phonological information. If this were true, the notion of advance planning of sentence intonation would need to assume a very large scope of phonological encoding that corresponds with the whole sentence or intonation phrase. This assumption goes against the original proposal (Levelt, 1989) that the phonological encoding entails the incremental retrieval of syllables or prosodic feet (i.e., a stressed syllable together with preceding or subsequent unstressed syllables). In other words, the length-dependent scaling of sentence-initial intonation peaks is highly unlikely to depend on the output of the phonological encoding because according to the model of lexical incrementality (among many others, see Levelt, 1989; Levelt et al., 1991), the syllable scores for the last-mentioned noun phrases are retrieved just before the articulation, that is, close to the end of an utterance.

However, a large scope of phonological planning is not that implausible; some earlier studies have reported non-incremental phonological encoding of transitive sentences in English and German (see Ferreira and Swets (2002); Oppermann et al. (2010), respectively). This means, there is evidence that transitive sentences with a subject-verb-object ordering (SVO sentences) are simple enough to be encoded phonologically to the end of the sentence (all word forms would be available already before the onset of speech). They should also be phonologically simple enough to be buffered in the working memory until the articulation processes are initiated. The relatively long speech onset latencies support the possibility that speakers of the current study planned with a large phonological planning scope. However, the increasing number of prosodic breaks in longer utterances argues against this possibility (for pauses as a signal for more incremental planning processes, see, e.g., Swets et al. 2021). Therefore, the degree of phonological lookahead remains an open question that certainly deserves further studies using more controlled methodologies and a variety of sentence structures.

On the other hand, a modified noun phrase represents a syntactic dependency relationship between the head noun and its modifier. The results also indicate that the tonal scaling of the sentence-initial intonation peaks might have been sensitive to the syntactic complexity in the modified noun phrases, and the question is how the sentence planning processes might have detected and stored the syntactic complexity of the last-mentioned constituent. The concrete content of linguistic representation at the stage of grammatical encoding is still an open matter (see, e.g., Branigan & Pickering, 2017). According to the lexical account, the linguistic representation might constitute itself as an interface between the levels of syntax and semantics, where a verb lemma and its valency determine the sentence form—i.e., the number of arguments a verb can semantically take (Bock & Levelt, 1994; Momma & Ferreira, 2019; Momma, Slevc, & Phillips, 2016; Sauppe, 2017). For sentence planning this means that the syntactic form of an utterance would emerge (incrementally) together with the retrieval of the verb (Bock & Levelt, 1994; Levelt, 1989; Momma & Ferreira, 2019; Momma et al., 2016). Access to the verb lemma would reveal the required syntactic functions. The following process of grammatical encoding would provide the correct order of these functions and initiate the incremental retrieval of the corresponding noun lemmas (see, e.g., Levelt, 1989). An alternative abstract syntactic account posits that the linguistic representation could also constitute an abstraction of syntactic relationships in the form of a sentence or dependency tree (Griffin & Bock, 2000; Kuchinsky et al., 2011). The current study and the data do not aim to settle the debate on the linguistic representation of sentence production. However, the results require a discussion about which of these representations could guide the advance planning of sentence intonation.

Notably, when the distractors were included in the pictures, the speakers in the current study gazed for a longer proportion of the speech onset latency at entities other than the agent. Possibly, during the earliest processing window, the conceptualization processes identified the visual contrast between the patient and the distractor and passed it on to the grammatical encoding processes in a form that currently remains unclear. According to the abstract syntactic account, the grammatical encoding processes might have generated a more complex dependency structure for the noun phrase within the syntactic structure of the entire incipient utterance. In line with the lexical account, it is also possible that the formulator assembled an abstract dependency relationship only for the single constituent (i.e., the patient) and buffered this whilst the rest of the syntactic structure emerged incrementally together with the verb.

With regard to intonation planning in the linguistically complex speaking situations, speakers might have expected a greater number of syntactic dependencies: (i) the one that is defined by the verb (ii) and another that is defined by the grammatical function of the object. Subsequently, the syntactic dependency relationships might have been picked up by the prosodic structure, within which the parameters of the F0 declination (of which the height of utterance-initial or phrase-initial intonation peaks is a part) might have been set (for the prosodic structure controlling for the F0 declination, see Bruce, 1977; Cooper & Sorensen, 1981; Féry & Truckenbrodt, 2005; Truckenbrodt, 2002). Accordingly, the results may speak for the idea that the intonation planning involves the generation of a prosodic structure that refers to syntactic structure (Nespor & Vogel, 1986; Selkirk, 1980, 1981, 1986, 2011) or in terms of psycholinguistics, to the higher-level planning processes such as grammatical encoding (Keating & Shattuck-Hufnagel, 2002;

Krivokapić, 2012; Krivokapić, 2007). So the linguistic representation to which the intonational planning might refer contains an abstract representation of concepts mapped onto the syntactic functions, their hierarchical relationships and, consequently, the correct ordering of semantic arguments of an incipient utterance.

Alternatively to the syntactic representation, some research suggests that the activation of concepts involves some hierarchical structuring already at the level of message planning (Bock & Ferreira, 2014; Griffin & Bock, 2000; Sauppe et al., 2013). Thus, the proposal is that the concepts at the message level emerge together with some hierarchical structure, which then guides the incremental process of lexical/grammatical encoding. In this account, the model for sentence production would be somewhat simpler. The speech production system would proceed from the hierarchical conceptual encoding processes straight to the incremental lemma retrieval and phonological encoding. The analyses of time courses of sentence production in Appendix 1 might lend support to this view. In particular, it can be observed from **Figures 7** and **8** that the timecourses for the two experiments in the time frame of 800 ms after picture presentation differ for the two experiments. In particular, the requirement to mention an attribute results in significant differences between the conditions with the separable attribute absent and present (see **Figure 7b** and **7c**). In contrast, no effect of conceptual complexity (the factor Separable Attribute) occurred in Experiment II (see **Figure 8b** and **8c**), most probably because the mentioning of the attributes was not required. The requirement to mention an attribute constituted a structural condition that effectively influenced the stage of conceptual planning in Experiment I. Therefore, the eye movements may indicate that in Estonian, the early event apprehension might have smoothly transitioned into the hierarchical encoding processes, in line with the theory of hierarchical incrementality (for hierarchical incrementality, see, e.g., Bock et al., 2004; Griffin & Bock, 2000; Kuchinsky et al., 2011; Lashley, 1951; Rosenbaum, Cohen, Jax, Weiss, & van der Wel, 2007; Wheeldon et al., 2013).

Consequently, intonation planning might refer to some conceptual hierarchy instead of syntactic representation. If so, then intonation planning might not involve the generation of the prosodic structure, given that the prosodic structure needs to refer to the syntax of utterances. Alternatively, a syntax-free prosodic hierarchy might be generated based on just the conceptual hierarchy. In this case, the components of prosodic hierarchy would reflect the hierarchical organization of the concepts activated by the process of message generation and not the boundaries and dependencies of syntactic constituents.

Another possibility is that the parameter of the utterance-initial intonation peaks is just set based on the number of concepts and the dimensions they involve. For example, in the linguistically complex speaking situations of the current study, the concept of the patient involved the dimension of attributes (e.g., colours). Thus, the intonation planning would only need to detect the more detailed conceptual space and set the declination parameter globally for an entire utterance. In this way, the advance planning of sentence intonation can be made to fit with the model of the incremental emergence of sentence intonation in the cognitive model of speaking (in Levelt, 1989). The declination parameter would be set together with the key and register based on the conceptual planning processes, and the rest of the prosodic structure would arise incrementally during the articulation. The problem with the notion of language-free selection of the declination parameter is that this idea ignores the evidence that F0 declination forms a part of the prosodic structure (Bruce, 1977; Féry & Ishihara, 2009; Féry & Truckenbrodt, 2005; Truckenbrodt, 2002). Perhaps the planning of sentence intonation still requires the assumption of the generation of prosodic structure (generated based on the syntactic structure) or prosodic hierarchy (generated based on the conceptual hierarchy). The current study, however, leaves the matter of the generation of the abstract prosodic representation open. A more direct test for the early emergence of prosodic structure would involve a speech production study where speakers produce actor names with varying syllable numbers (e.g., TV vs. television). This test would eliminate the impact of syntactic complexity and investigate whether the intonation planning relies on the early activation of prosodic frames but still leaves open the order of the activation of syntactic and prosodic structures (either prosody first or syntax first).

More importantly, the results of the current study provide a generalization that the advance planning of sentence intonation is preconditioned by the large increment of conceptual planning. In other words, the length-dependent scaling of the sentence-initial intonation peaks might be possible only in situations where the linguistic encoding is preceded by the generation of the comprehensive relational framework of incipient utterances. In this regard, Estonian sentence planning, with its overall preference for the less incremental conceptual planning strategy, constitutes a convenience for the advance planning of sentence intonation. The earlier research has identified a link between the non-incremental conceptual planning and morphological case-marking (Norcliffe et al., 2015; Sauppe et al., 2013). The findings of this study corroborate this link and support the prediction that the advance planning of sentence intonation should occur in other case-morphological languages with a similar regularity to that at which it occurs in Estonian (in addition to this study, see also Asu et al., 2017, 2016). Conversely, it should be more difficult to find advance planning of sentence intonation in languages that prefer a more incremental planning strategy to a non-incremental planning strategy for the stage of conceptualization. The speed at which speakers initiate speech could serve as an additional test condition to investigate the relationship between non-incremental conceptual planning and the advance planning of sentence intonation. When asked to start speaking quickly, speakers may have less opportunity to plan in advance compared to situations with more flexible time constraints. Thus, the advance planning of sentence intonation might also be more difficult in situations of fast speech.

## 5. Conclusion

To conclude, the study presents a novel synthesis between the phonetic account of advance planning of sentence intonation and the psycho-linguistic account of sentence planning. The two visual world speech production experiments demonstrated the importance of the rapid and non-incremental generation of the conceptual framework for the planning of sentence intonation in Estonian. In particular, a detailed and possibly structurally informed hierarchical conceptual framework was argued to guide the intonation planning for the production of spontaneous speech. Moreover, the study corroborated the link between the hierarchically incremental sentence planning and the case-morphological language type. In general, the results of the study establish advance planning of sentence intonation as an integral part of the processes of language planning and production and motivate the incorporation of the phonetic findings—concrete intonational phenomena in particular—into the psycho-linguistic study of sentence planning.

# Appendices

## Appendix 1

**Figures 7d** and **8d** demonstrate the typical analyses of averaged proportions of eye fixations separated into 20-millisecond time bins. The plots above (a–c) demonstrate the time frames of significant differences as acquired by the generalized additive mixed-effects modelling (GAMMs) implemented in the *mgcv* package in R (Wood 2004, 2017). For this, the Area of Interest (AOI) was defined as a dependent ordinal numeric variable biased towards the agent (agent > distractor > attribute > patient). The GAM models strictly followed the methodology presented in Wieling (2018) and assessed the likelihood of fixating on one of the four AOIs as a function of time. **Figures 7** and **8** demonstrate the time frames of significant differences between the presence and absence of a distractor (A; see **Figure 1** for distractors and attributes) and between the presence and absence of a separable attribute (SA) in the absence of distractor (B) and in the presence of a distractor (C). The smaller the differences (below the zero-line), the likelier the fixations on the agents in a particular time bin.

## Appendix 2

| | ***RE:* . ~ … + (1\|sub) + (1\|item) + (1\|trial)** | |
|---|---|---|
| | **Estimate** | **CI** |
| Intercept | 0.019 | –0.086–0.124 |
| Duration (ms) | 0.118$^{***}$ | 0.086–0.150 |
| Experiment | 0.095 | –0.0003–0.191 |
| Duration:Experiment | –0.011 | –0.042–0.019 |
| Observations | 3,989 | |

**Table 3:** The parameters of the fixed effects and their confidence intervals (CI) as a result of the LMER fitting. F0 maxima (in semitones) were modelled as a function of an interaction between the utterance duration and Experiment. The probabilities of the model parameters were obtained with the help of the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017), and these provide certainty that a particular parameter value will be observed in some future data. At the top: the full random effects structure.
*Note:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

| | ***RE*: . ~ ... + (1+attribute\|sub) + (1\|item) + (1\|trial)** | |
|---|---|---|
| | **Estimate** | **CI** |
| Intercept | 0.390$^{***}$ | 0.376 — 0.404 |
| Distractor | 0.027$^{***}$ | 0.022 — 0.033 |
| S. Attribute | 0.013$^{***}$ | 0.007 — 0.019 |
| Experiment | –0.002 | –0.009 — 0.005 |
| Distractor:S. Attribute | 0.004 | –0.002 — 0.009 |
| Distractor:Experiment | 0.005 | –0.0003 — 0.011 |
| S. Attribute: Experiment | –0.001 | –0.007 — 0.004 |
| Distractor:S. Attribute:Experiment | 0.005 | –0.0003 — 0.010 |
| Observations | | 4,205 |

**Table 4:** The parameters of the fixed effects and their confidence intervals (CI) as a result of the LMER fitting. The intercept represents the mean across all the conditions, and the parameters estimate the deviations from the mean. The relative duration of the naming gaze (%) was modelled as a function of an interaction between Distractor, Separable Attribute and Experiment. The probabilities of the model parameters were optained with the help of the *lmerTest* package (Kuznetsova et al., 2017), and these provide certainty that a particular parameter value will be observed in some future data. At the top: the random effects structure (RE) that converged.
*Note:* $^{*}$p < 0.05; $^{**}$p < 0.01; $^{***}$p < 0.001.

| | ***RE*: . ~ ... + (1\|sub) + (1\|item) + (1\|trial)** | |
|---|---|---|
| | **Estimate** | **CI** |
| Intercept | 0.029 | –0.078 — 0.137 |
| Rel. gaze dur. | 0.003 | –0.025 — 0.032 |
| Experiment | 0.123$^{*}$ | 0.026 — 0.220 |
| Rel. gaze dur.:Experiment | –0.015 | –0.044 — 0.013 |
| Observations | | 3,989 |

**Table 5:** The parameters of the fixed effects and their confidence intervals (CI) as a result of the LMER fitting. F0 maxima (in semitones) were modelled as a function of an interaction between the relative duration of the naming gaze and Experiment. The probabilities of the model parameters were obtained with the help of the *lmerTest* package (Kuznetsova et al., 2017), and these provide certainty that a particular parameter value will be observed in some future data. At the top: the full random effects structure.
*Note:* $^{*}$p < 0.05; $^{**}$p < 0.01; $^{***}$p < 0.001.

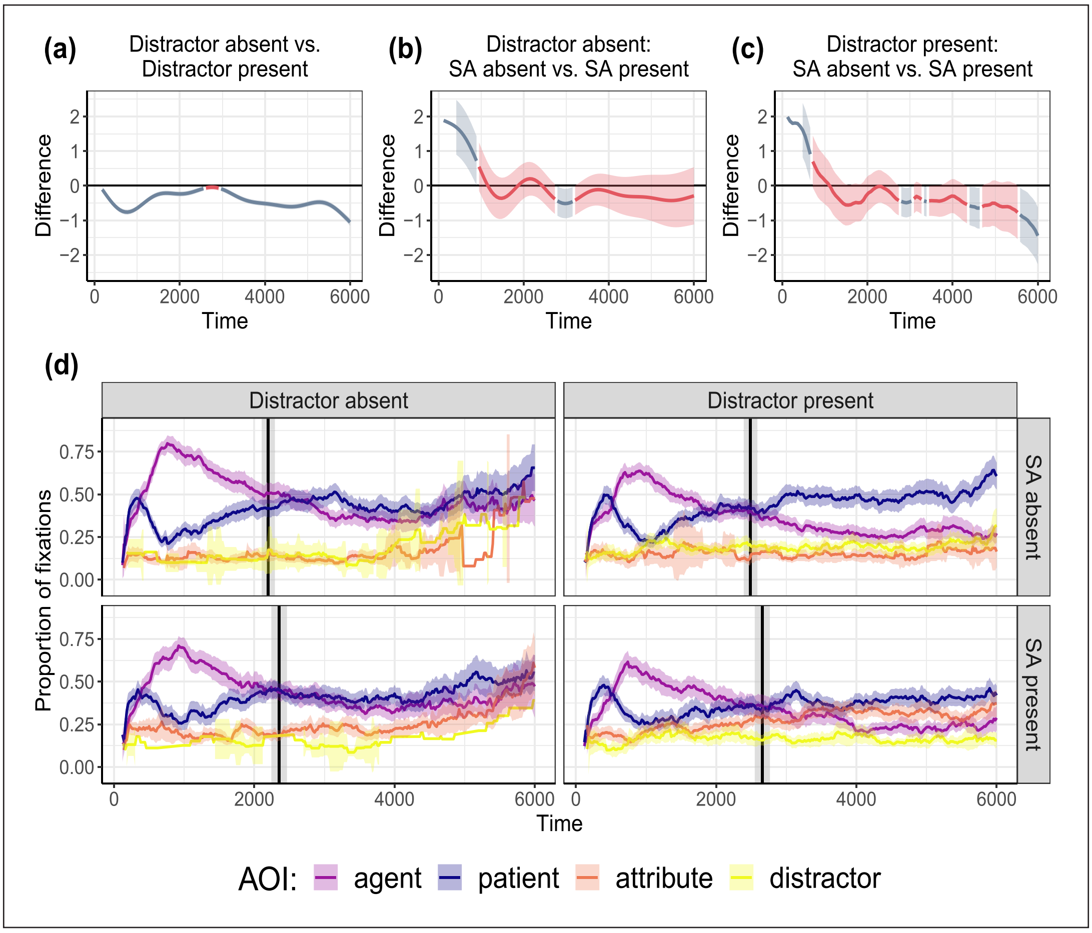

**Figure 7:** (a), (b) and (c): Functions of significant differences between the conditions as obtained from the generalized additive modelling of eye gaze (SA refers to Separable Attribute). The difference between the conditions is significant when the function and its confidence interval (shaded area around it) do not overlap with the horizontal line at zero (grey indicates significant differences). (d): Averaged time courses of the production of short and long sentences containing bare nouns (e.g., *Mees sikutab eeslit* 'A man is pulling a donkey') nouns modified with attributive adjectives (e.g., *Mees sikutab suurt eeslit* 'A man is pulling a big donkey'), and nouns modified with attributive nouns (e.g., *Mees sikutab korviga eeslit* 'A man is pulling a donkey with a basket'). The averaged proportions of fixations directed towards agents are shown with purple lines, the fixations towards patients are indicated with blue lines, and the averaged proportions of fixations directed towards attributes and distractors are demonstrated with the pink and yellow lines, respectively. The vertical black lines indicate the means of speech onset latencies. The shaded areas around the functions and lines indicate the 95% confidence intervals.

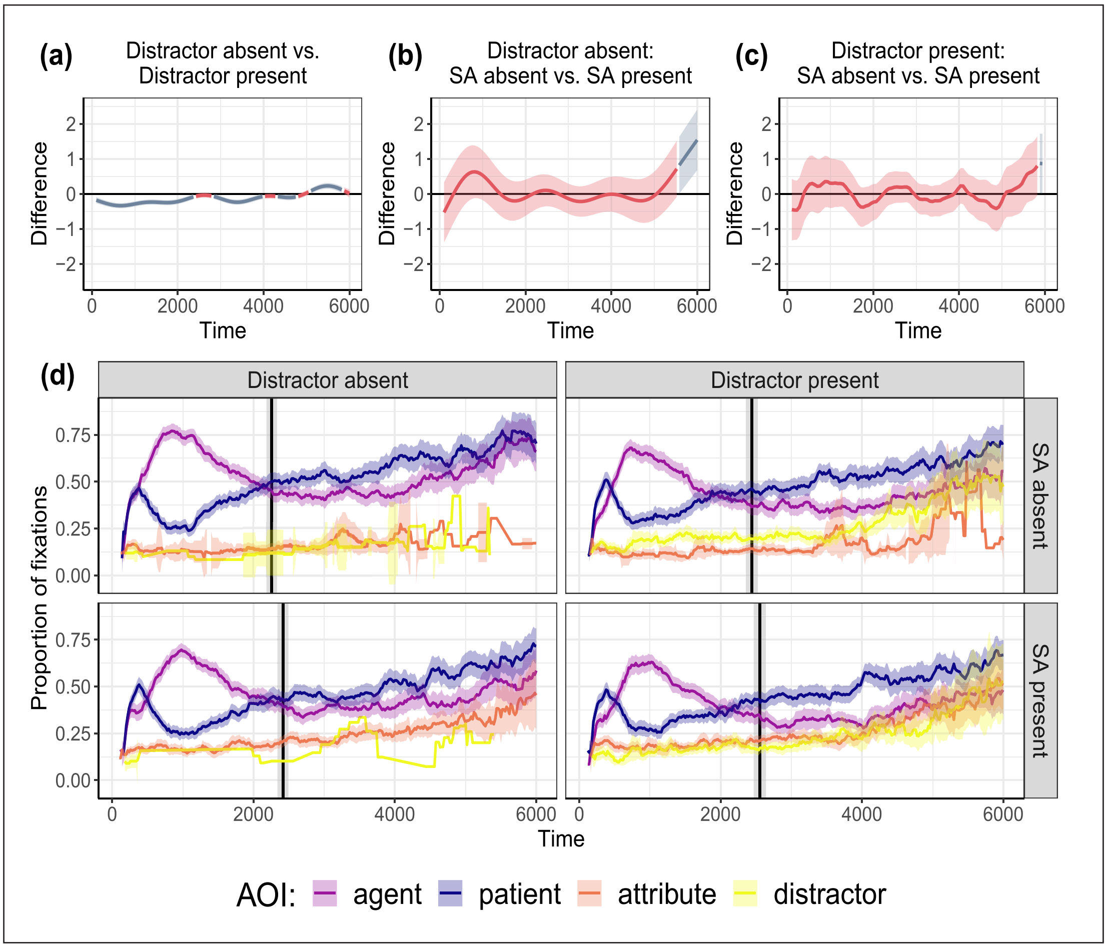

**Figure 8:** (a), (b) and (c): Functions of significant differences between the conditions as obtained from the generalized additive modelling of eye gaze (SA refers to Separable Attribute). The difference between the conditions is significant when the function and its confidence interval (shaded area around it) do not overlap with the horizontal line at zero (grey indicates significant differences). (d): Averaged time courses of the production of short sentences containing bare nouns (e.g., *Mees sikutab eeslit* 'A man is pulling a donkey') across the four conditions of varying conceptual complexity. The averaged proportions of fixations directed towards agents are shown with purple lines, the fixations towards patients are indicated by blue lines, and the averaged proportions of fixations directed towards attributes and distractors by pink and yellow lines, respectively. The vertical black lines indicate the means of speech onset latencies. The shaded areas around the functions and lines indicate the 95% confidence intervals.

## Appendix 3

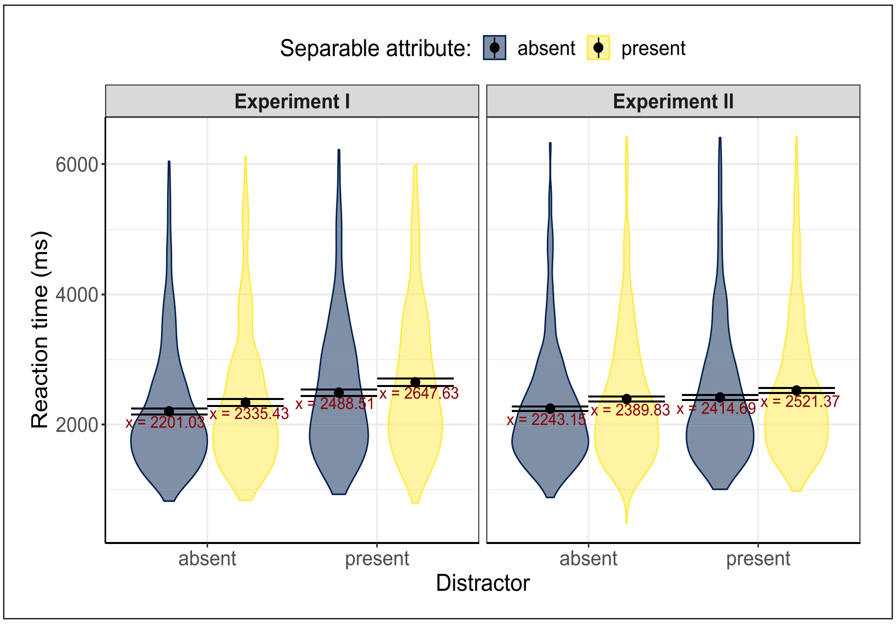

**Figure 9:** The distributions of the speech onset latencies across the two experiments as a function of the factors Distractor (absent vs. present) and Separable Attribute (absent vs. present). The condition in which a separable attribute and a distractor are absent (the grey colour indicates the absence and the yellow colour the presence of separable attributes) is the simplest, containing just the two interacting actors, whereas the condition with a separable attribute and a distractor present is the most complex, containing three actors and an object as a characteristic feature of the interacting patient (see **Figure 1**). The black labelled circles indicate the means of the variables and the error bars represent the 95% confidence intervals. The LMER analysis reports significant differences as long as the 95% confidence intervals do not overlap.

## Abbreviations

The following abbreviations were used in the interlinear gloss in Section (2.4):

SG.NOM: singular nominative case

SG.COM: singular comitative case (translating 'with' in the source)

3.SG.PRS: third person, singular, present

## Additional files

The data together with the analysis scripts are available in the OSF repository https://osf.io/87a59/. Speech recordings can be made available from the corresponding author on reasonable request.

## Acknowledgements

This study would not exist without the great colleagues and friends who have helped me on the path of this scientific quest. Firstly, I would like to express my deep gratitude to Agnieszka Konopka, who contributed to the grant writing, conceptualization, design and materials of the current study. Secondly, I am extremely thankful to Kairi Kreegipuu and Tiit Mogom of the Institute of Psychology at the University of Tartu, who provided access to the eye-tracking system and assisted in setting up the first experiment. Thirdly, a special thanks belongs to Markus Bader, Helen Hint, Frank Kügler, Corinna Langer, Yvonne Portele, Maria Reile, Piia Taremaa and Roger Yallop for their highly valuable feedback, support and motivating comments on the numerous previous versions of this article. Furthermore, the running of the second experiment succeeded with the help of the excellent research assistants Mari Aigro and Mariann Proos (recruiting and recording of the participants), Kerstin Sillaots (transcription of the utterances) and Annika Pant (manual check of the forced alignment of word boundaries in the materials). Finally, I am deeply grateful to the acting editor, Bettina Braun, for her consistently constructive and helpful critiques throughout the multiple revisions of the submitted manuscript.

## Funding information

Post-doctoral research grant from the Fritz Thyssen Stiftung ("Planning sentences and sentence intonation cross-linguistically," 10.18.2.040SL).

## Competing interests

The author has no competing interests to declare.

## References

Asu, E. L. (2004). *The phonetics and phonology of Estonian intonation* (Doctoral dissertation). University of Cambridge.

Asu, E. L. (2005). Towards a phonological model of Estonian intonation. In *Proceedings of the second Baltic conference on human language technologies, Tallinn 4–5 May 2005* (pp. 95–100).

Asu, E. L., Lippus, P., Sahkai, H., & Salveste, N. (2017). F0 declination in read vs. spontaneous Estonian. In *Nordic prosody: Proceedings of the XIIth conference, Trondheim 2016* (pp. 63–72). DOI: https://doi.org/10.3726/b11152

Asu, E. L., Lippus, P., Salveste, N., & Sahkai, H. (2016). F0 declination in spontaneous Estonian: Implications for pitch-related preplanning. In *Proceedings of Speech Prosody, Boston 31 May – 3 June 2016*. DOI: https://doi.org/10.21437/SpeechProsody.2016-234

Asu, E. L., & Nolan, F. (1999). The effect of intonation on pitch cues to the Estonian quantity contrast. In *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 1873–1876).

Atkinson, J. E. (1978). Correlation analysis of the physiological factors controlling fundamental voice frequency. *The Journal of the Acoustical Society of America*, *63*(1), 211–222. Retrieved from https://asa.scitation.org/doi/abs/10.1121/1.381716. DOI: https://doi.org/10.1121/1.381716

Badecker, W., & Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of Memory and Language*, *56*(1), 65–85. Retrieved from https://www.sciencedirect.com/science/article/pii/S0749596X06001124. DOI: https://doi.org/10.1016/j.jml.2006.08.004

Bock, K., & Ferreira, V. S. (2014). Syntactically speaking. In *The Oxford handbook of language* (pp. 21–46). Oxford University Press.

Bock, K., Irwin, D. E., & Davidson, D. J. (2004). Putting first things first. In J. Henderson, & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 249–278). New York, NY, US: Psychology Press.

Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). Academic Press.

Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer (Version 6.1.09)*. Retrieved from http://www.praat.org/

Bois, J. W. D., Cumming, S., Schuetze-Coburn, S., & Paolino, D. (Eds.). (1992). *Discourse transcription* (Vol. 4). Department of Linguistics, University of California, Santa Barbara.

Branigan, H. P., & Pickering, M. J. (2017). An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, *40*, e282. DOI: https://doi.org/10.1017/S0140525X16002028

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and cognitive processes*, *25*(7), 1044–1098. DOI: https://doi.org/10.1080/01690965.2010.504378

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Geerups.

Buhmann, J., Caspers, J., van Heuven, V. J., Hoekstra, H., Martens, J.-P., & Swerts, M. (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the spoken Dutch corpus. In *Proceedings of the third international conference on language resources and evaluation (LREC'02)*. Retrieved from http://www.lrec-conf.org/proceedings/lrec2002/pdf/96.pdf

Cole, J. R., & Reitter, D. (2019). The role of working memory in syntactic sentence realization: A modeling & simulation approach. *Cognitive Systems Research*, *55*, 95–106. Retrieved from http://www.sciencedirect.com/science/article/pii/S138904171830353X. DOI: https://doi.org/10.1016/j.cogsys.2019.01.001

Cooper, W. E., & Sorensen, J. M. (1981). *Fundamental frequency in sentence production*. Springer-Verlag. DOI: https://doi.org/10.1007/978-1-4613-8093-1

Duchowski, A. (2007). *Eye tracking methodology: Theory and practice*. Springer.

Elordieta, G., & Selkirk, E. (2022). Unaccentedness and the formation of prosodic structure in Lekeitio Basque. In *Prosody and Prosodic Interfaces*. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198869740.003.0013

Elsner, M., Clarke, A., & Rohde, H. (2018). Visual complexity and its effects on referring expression generation. *Cognitive Science*, *42*(S4), 940–973. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12507. DOI: https://doi.org/10.1111/cogs.12507

Erelt, M. (2003). Preface. Estonian language. In *Linguistica Uralica: Supplementary series* (Vol. 1, p. 7–8). Estonian Academy of Sciences.

Erelt, M., & Metslang, H. (Eds.). (2017). *Eesti keele süntaks [eng. Estonian syntax]* (No. 3). Tartu Ülikooli Kirjastus.

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*, 57–84. DOI: https://doi.org/10.1006/jmla.2001.2797

Féry, C., & Ishihara, S. (2009). How focus and givenness shape prosody. In M. Zimmermann, & C. Féry (Eds.), *Information structure: Theoretical, typological, and experimental perspectives* (pp. 36–63). Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199570959.003.0003

Féry, C., & Truckenbrodt, H. (2005). Sisterhood and tonal scaling. *Studia Linguistica*, *59*, 223–243. DOI: https://doi.org/10.1111/j.1467-9582.2005.00127.x

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, *47*(1), 27–52. DOI: https://doi.org/10.2307/412187

Fuchs, S., Petrone, C., Krivokapić, J., & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, *41*, 29–47. DOI: https://doi.org/10.1016/j.wocn.2012.08.007

Garrett, M. F. (1975). The analysis of sentence production. In G. H. Bower (Ed.), (Vol. 9, pp. 133–177). Academic Press. DOI: https://doi.org/10.1016/S0079-7421(08)60270-4

Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language production* (pp. 177–220). Academic Press.

Gatt, A., Krahmer, E., van Deemter, K., & van Gompel, R. P. (2017). Reference production as search: The impact of domain size on the production of distinguishing descriptions. *Cognitive Science*, *41*(S6), 1457–1492. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12375. DOI: https://doi.org/10.1111/cogs.12375

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*(544–596). DOI: https://doi.org/10.1016/j.jml.2007.01.007

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, *82*, B1–B14. DOI: https://doi.org/10.1016/S0010-0277(01)00138-X

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*, 274–279. DOI: https://doi.org/10.1111/1467-9280.00255

Griffin, Z. M., & Davison, J. C. (2011). A technical introduction to using speakers' eye movements to study language. In G. Jarema, G. Libben, & C. Westbury (Eds.), *Methodological and analytic frontiers in lexical research (part ii)* (pp. 53–82). DOI: https://doi.org/10.1075/ml.6.1.03gri

Hartsuiker, R. J., & Barkhuysen, P. N. (2006). Language production and working memory: The case of subject-verb agreement. *Language and Cognitive Processes*, *21*(1–3), 181–204. DOI: https://doi.org/10.1080/01690960400002117

Himmelmann, N. P. (2022). Prosodic phrasing and the emergence of phrase structure. *Linguistics*, *60*(3), 715–743. DOI: https://doi.org/10.1515/ling-2020-0135

Himmelmann, N. P., Sandler, M., Strunk, J., & Unterladstetter, V. (2018). On the universality of intonational phrases: a cross-linguistic interrater study. *Phonology*, *35*(2), 207–245. DOI: https://doi.org/10.1017/S0952675718000039

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking. A comprehensive guide to methods and measures*. Oxford University Press.

Honda, K. (2004). Physiological factors causing tonal characteristics of speech: From global to local prosody. In *Proceedings Speech Prosody 2004* (pp. 739–744). DOI: https://doi.org/10.21437/SpeechProsody.2004-171

Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in F0 control. *Language and Speech*, *42*(4), 401–411. (PMID: 10845244). DOI: https://doi.org/10.1177/00238309990420040301

Hwang, H., & Kaiser, E. (2015). Accessibility effects on production vary cross-linguistically: Evidence from English and Korean. *Journal of Memory and Language*, *84*, 190–204. DOI: https://doi.org/10.1016/j.jml.2015.06.004

Kaisse, E. M. (1985). *Connected speech: The interaction of syntax and phonology*. Academic Press.

Keating, P. A. (2006). Phonetic encoding of prosodic structure. In J. Harrington, & M. Tabain (Eds.), *Speech production: Models, phonetic processes and techniques* (pp. 167–186). Psychology Press.

Keating, P. A., & Shattuck-Hufnagel, S. (2002, August). A prosodic view of word form encoding for speech production. *UCLA Working Papers in Phonetics*, *101*, 112–156.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. DOI: https://doi.org/10.1016/j.csl.2017.01.005

Konopka, A. E., & Meyer, A. S. (2014). Priming sentence planning. *Cognitive Psychology*, *73*, 1–40. DOI: https://doi.org/10.1016/j.cogpsych.2014.04.001

Kratzer, A., & Selkirk, E. (2020). Deconstructing information structure. *Glossa: a journal of general linguistics*, *5*(1), 113. DOI: https://doi.org/10.5334/gjgl.968

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, *35*(2), 162–179. Retrieved from http://www.sciencedirect.com/science/article/pii/S0095447006000180. DOI: https://doi.org/10.1016/j.wocn.2006.04.001

Krivokapić, J. (2012). Prosodic planning in speech production. In S. Fuchs, M. Weirich, D. Pape, & P. Perrier (Eds.), *Speech planning and dynamics* (pp. 157–190). Bern, Switzerland: Peter Lang. Retrieved from https://www.peterlang.com/view/9783653014389/9783653014389.00008.xml

Kuchinsky, S. E., Bock, K., & Irwin, D. E. (2011). Reversing the hands of time: Changing the mapping from seeing to saying. *Journal of experimental psychology: Learning, memory, and cognition*, *37*(3), 748–756. DOI: https://doi.org/10.1037/a0022637

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. DOI: https://doi.org/10.18637/jss.v082.i13

Lashley, K. S. (1951). The problem of serial order in behavior. In *Cerebral mechanisms in behavior; the Hixon Symposium* (p. 112–146). Wiley.

Lee, S. J., & Selkirk, E. (2022). Xitsonga tone: The syntax–phonology interface. In *Prosody and Prosodic Interfaces.* Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198869740.003.0012

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

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *The behavorial and brain sciences*, *22*(1), 1–38; discussion 38–75. DOI: https://doi.org/10.1017/S0140525X99001776

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, *98*, 122–142. DOI: https://doi.org/10.1037/0033-295X.98.1.122

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff, & R. T. Oehrle (Eds.), *Language sound structure* (pp. 155–233). MIT Press.

Meyer, A. S., Roelofs, A., & Brehm, L. (2019). Thirty years of speaking: An introduction to the special issue. *Language, Cognition and Neuroscience*, *34*(9), 1073–1084. DOI: https://doi.org/10.1080/23273798.2019.1652763

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Chapman and Hall/CRC.

Momma, S., & Ferreira, V. S. (2019). Beyond linear order: The role of argument structure in speaking. *Cognitive Psychology*, *114*, 101228. Retrieved from https://www.sciencedirect.com/science/article/pii/S001002851930218X. DOI: https://doi.org/10.1016/j.cogpsych.2019.101228

Momma, S., Slevc, L. R., & Phillips, C. (2016). The timing of verb selection in Japanese sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 813–824. DOI: https://doi.org/10.1037/xlm0000195

Myachykov, A., & Garrod, S. (2008). Perception and word order in Russian and Finnish sentence production. In W. Ramm, & C. Fabricius-Hansen (Eds.), *Linearisation and segmentation in discourse: Multidisciplinary approaches to discourse.* Deptartment of Literature, Area Studies and European Languages, University of Oslo.

Myachykov, A., Scheepers, C., Garrod, S., Thompson, D., & Fedorova, O. (2013). Syntactic flexibility and competition in sentence production: The case of English and Russian. *Quarterly Journal of Experimental Psychology*, *66*(8), 1601–1619. (PMID: 23286507). DOI: https://doi.org/10.1080/17470218.2012.754910

Myachykov, A., & Tomlin, R. S. (2008). Perceptual priming and structural choice in Russian sentence production. *Journal of Cognitive Science*, *6*(1), 31–48. DOI: https://doi.org/10.17791/jcs.2008.9.1.31

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Foris.

Norcliffe, E., Konopka, A. E., Brown, P., & Levinson, S. C. (2015). Word order affects the time course of sentence formulation in Tzeltal. *Language, Cognition and Neuroscience*, *30*(9), 1187–1208. DOI: https://doi.org/10.1080/23273798.2015.1006238

Odden, D. A. (1987). Kimatuumbi phrasal phonology. *Phonology Yearbook*, *4*, 13–26. DOI: https://doi.org/10.1017/S0952675700000750

Odden, D. A. (1990). Syntax, lexical rules and postlexical rules in Kimatuumbi. In S. Inkelas, & D. Zec (Eds.), *The phonology-syntax connection* (pp. 259–277). University of Chicago Press.

Oppermann, F., Jescheniak, J. D., & Schriefers, H. (2010). Phonological advance planning in sentence production. *Journal of Memory and Language*, *63*(4), 526–540. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X10000628. DOI: https://doi.org/10.1016/j.jml.2010.07.004

Prieto, P., D'Imperio, M., Elordieta, G., Frota, S., & Vigário, M. (2006). Evidence for 'soft' preplanning in tonal production: Initial scaling in Romance. In *Speech Prosody, Dresden 2–5 May 2006* (pp. 803–806). DOI: https://doi.org/10.21437/SpeechProsody.2006-169

Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & van der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Human Movement Science*, *26*(4), 525–554. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167945707000280 (European Workshop on Movement Science 2007). DOI: https://doi.org/10.1016/j.humov.2007.04.001

Sauppe, S. (2017). Word order and voice influence the timing of verb planning in German sentence production. *Frontiers in Psychology*, *8*, 1648. Retrieved from https://www.frontiersin.org/article/10.3389/fpsyg.2017.01648. DOI: https://doi.org/10.3389/fpsyg.2017.01648

Sauppe, S., Norcliffe, E., Konopka, A. E., Valin, R. D. J. V., & Levinson, S. C. (2013). Dependencies first: Eye tracking evidence from sentence production in Tagalog. In *Proceedings of the 35th annual meeting of the Cognitive Science Society (CogSci 2013)* (pp. 1265–1270).

Selkirk, E. (1980). Prosodic domains in phonology: Sanskrit revisited. In M. Aronoff, & R. T. Oehrle (Eds.), *Language sound structure* (pp. 107–136). Cambridge, MA: MIT Press.

Selkirk, E. (1981). On prosodic structure and its relation to syntactic structure. In T. Fretheim (Ed.), *Nordic prosody* (pp. 111–140). TAPIR.

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook*, *3*, 371–405. DOI: https://doi.org/10.1017/S0952675700000695

Selkirk, E. (2011). The syntax-phonology interface. In J. Goldsmith, J. Riggle, & A. Yu (Eds.), *The handbook of phonological theory* (2nd ed., pp. 435–484). DOI: https://doi.org/10.1002/9781444343069.ch14

Shattuck-Hufnagel, S. (2019). Toward an (even) more comprehensive model of speech production planning. *Language, Cognition and Neuroscience*, *34*(9), 1202–1213. DOI: https://doi.org/10.1080/23273798.2019.1650944

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2022). Afex: Analysis of factorial experiments [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=afex (R package version 1.3-1)

Slevc, L. R. (2011). Saying what's on your mind: Working memory effects on sentence production. *Journal of experimental psychology: Learning, memory, and cognition*, *37*(6), 1503–14. DOI: https://doi.org/10.1037/a0024350

Speekenbrink, M. (2022). *Statistics: Data analysis and modelling.* Retrieved 21. September 2023, from https://mspeekenbrink.github.io/sdam-book/ (Book published on personal Website)

Strik, H., & Boves, L. (1992). Control of fundamental frequency, intensity and voice quality in speech. *Journal of Phonetics*, *20*(1), 15–25. Retrieved from http://www.sciencedirect.com/science/article/pii/S0095447019302505. DOI: https://doi.org/10.1016/S0095-4470(19)30250-5

Swets, B., Fuchs, S., Krivokapić, J., & Petrone, C. (2021). A cross-linguistic study of individual differences in speech planning. *Frontiers in psychology*, *12*, 655516. DOI: https://doi.org/10.3389/fpsyg.2021.655516

Thorsen, N. G. (1980). A study of the perception of sentence intonation—Evidence from Danish. *The Journal of the Acoustical Society of America*, *3*(67), 1014–1030. DOI: https://doi.org/10.1121/1.384069

Thorsen, N. G. (1985). Intonation and text in standard Danish. *The Journal of the Acoustical Society of America*, *77*(3), 1205–1216. DOI: https://doi.org/10.1121/1.392187

Tomlin, R. S. (1995). Focal attention, voice, and word order. In P. Dowing, & M. Noonan (Eds.), *Word order in discourse* (pp. 517–552). John Benjamins. DOI: https://doi.org/10.1075/tsl.30.18tom

Tøndering, J. (2011). Preplanning of intonation in spontaneous versus read aloud speech: Evidence from Danish. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 2010–2013).

Truckenbrodt, H. (2002). Upstep and embedded register levels. *Phonology*, *19*, 77–120. DOI: https://doi.org/10.1017/S095267570200427X

van de Velde, M., Meyer, A. S., & Konopka, A. E. (2014). Message formulation and structural assembly: Describing "easy" and "hard" events with preferred and dispreferred syntactic structures. *Journal of Memory and Language*, *71*(1), 124–144. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X13001101. DOI: https://doi.org/10.1016/j.jml.2013.11.001

Viitso, T.-R. (2003). Structure of the Estonian language: Phonology, morphology and word formation. In *Estonian language* (Vol. 1, p. 9–92). Estonian Academy of Sciences.

Vilkuna, M. (1998). Word order in European Uralic. In A. Siewierska (Ed.), *Constituent order in the languages of Europe* (pp. 173–233). Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110812206.173

Wheeldon, L., & Lahiri, A. (1997). Prosodic units in speech production. *Journal of Memory and Language*, *37*(3), 356–381. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X97925171. DOI: https://doi.org/10.1006/jmla.1997.2517

Wheeldon, L., Ohlson, N., Ashby, A., & Gator, S. (2013). Lexical availability and grammatical encoding scope during spoken sentence production. *Quarterly journal of experimental psychology (2006)*, *66*(8), 1653–73. DOI: https://doi.org/10.1080/17470218.2012.754913

Wheeldon, L., & Smith, M. (2003). Phrase structure priming: A shortlived effect. *Language and Cognitive Processes*, *18*(4), 431–442. DOI: https://doi.org/10.1080/01690960244000063

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of English. *Journal of Phonetics*, *70*, 86–116. Retrieved from http://www.sciencedirect.com/science/article/pii/S0095447017301377. DOI: https://doi.org/10.1016/j.wocn.2018.03.002

Winter, B. (2019). *Statistics for linguists: An introduction using R* (1st ed.). Routledge. DOI: https://doi.org/10.4324/9781315165547

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673–686. DOI: https://doi.org/10.1198/016214504000000980

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781315370279

Yuan, J., & Liberman, M. (2014). F0 declination in English and Mandarin broadcast news speech. *Speech Communication*, *65*, 67–74. DOI: https://doi.org/10.1016/j.specom.2014.06.001