Journal of Phonetics 103 (2024) 101297

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Use of segmental detail as a cue to prosodic structure in reference to information structure in German

Holger Mitterer[a], Sahyang Kim[b], Taehong Cho[c,*]

[a] Department of Cognitive Science, University of Malta, Msida, Malta
[b] Department of English Education, Hongik University, Seoul, Korea
[c] Hanyang Institute for Phonetics and Cognitive Science, Department of English Language and Literature, Hanyang University, Seoul, Korea

ARTICLE INFO

Article history:
Received 16 March 2023
Received in revised form 10 January 2024
Accepted 19 January 2024
Available online 23 February 2024

Keywords:
Prosody
Prosodic structure
Segmental details
Speech processing
German
Verum focus
Information structure
Mouse-tracking experiment

ABSTRACT

Listeners often make use of suprasegmental features to compute a prosodic structure and thereby infer an information structure. In this study, we ask whether listeners also use segmental details as a cue to the prosodic structure (and thus also the information structure) of an utterance. To this end, we examined the effects of segmental variation of German auxiliary haben (‘to have’)—i.e., hyperarticulated [habən], moderately reduced [habm], and strongly reduced [ham]. Three remotely accessed online mouse-tracking experiments were carried out by adapting the lab-based experimental paradigms used in Roettger and Franke (2019). They showed effects of pitch accent on the auxiliary haben, leading to the interpretation of an affirmative answer to a preceding question, thus anticipating an upcoming referent noun to be the same as the one given in the question (i.e., the verum focus effect). Experiment 1 adapted the design Roettger and Franke (2019) to an online setting. In Experiment 2, listeners were indeed found to make use of the segmental detail of the auxiliary haben, even in the absence of f0 (pitch accent) information—i.e., the hyperarticulated (full) form showed an effect similar to the pitch accented form, albeit smaller. In Experiment 3, we confirmed that the observed segmental effects were not simply due to learning that might have taken place during the experiment. Our results thus imply that the analysis of prosodic structure, which is often assumed to occur in parallel with the segmental analysis, must integrate segmental details that help to signal the prosodic structure.

1. Introduction

The saying “it’s not (only) what you say but how you say it” implies that every speech act contains multilayered information about what and how. Those layers can be mapped, loosely speaking, onto the segmental and suprasegmental layers of the speech signal (Lehiste, 1970). From the perspective of speech production, it may be useful to consider these layers as independent, as follows. On the one hand, the segmental layer contributes in large part to the 'what' component, as it determines the actual content of an utterance. It contains the phonological specifications of individual segments, which are combined to form words, and these words are further grouped to create an utterance. On the other hand, the suprasegmental layer is primarily responsible for shaping the 'how' component, and it can be influenced by various higher-order linguistic structures, including information structure, intonational phonology, and prominence distribution. It is also influenced by non-linguistic factors like paralinguistic features. Those distinctions, however, might not be straightforward to a listener because the segmental and suprasegmental (prosodic) aspects are convoluted in the speech signal (for a review, see McQueen & Dilley, 2020). It is then the task of the listener to comprehend the segmental and suprasegmental message intended by the speaker, even though they are not conveyed through fully independent channels, because both segmental and suprasegmental characteristics of speech may originate from the same type of information. Thus, in this context, by independence we refer to the listeners' ability to use segmental and suprasegmental cues independently, regardless of whether both types of cues signal different types of information (e.g., segmental versus phrasal phonology) or stem from one type of information (e.g., hyperarticulation or reduction associated with phrasal-level prominence).

While the notion of ‘independence’ might suggest that these cues operate in distinct ways, it does not imply that they are

* Corresponding author.
E-mail address: tcho@hanyang.ac.kr (T. Cho).

https://doi.org/10.1016/j.wocn.2024.101297
0095-4470/© 2024 Elsevier Ltd. All rights reserved.

always processed entirely separately, each referring to different types of information. Instead, they often function interdependently, collaboratively signaling one type of information. One type of such interdependence can be found in Salverda et al. (2003), who showed that listeners use syllable duration (i.e., a prosodic aspect of the speech signal) to determine whether a syllable is a one-syllable word (e.g., *ham*) or the first syllable of a two-syllable word (e.g., *hamster*). That is, although both *ham* and *hamster* share the same phonological specifications (the *what* component), the way they are realized in the suprasegmental dimension influences how the *what* component should be processed. Another type of interdependence is evident in Kim and Cho (2013). They showed that listeners take into account the prosodic boundary information embedded along the suprasegmental dimension to decide whether an upcoming voice onset time (VOT) is more likely to signal a voiced or voiceless stop in English (see also Mitterer et al., 2016 for related results in English; Steffman et al., 2022 in Korean). Kim et al. (2018) built on these studies and demonstrated that listeners exploit suprasegmental information to recover underlying segmental information from a surface form altered by the application of a phonological rule. All of these studies suggest that the manner in which an utterance is delivered (the 'how') plays a crucial role in interpreting the content of the utterance (the 'what').

The *how* component of an utterance, however, is expressed not only along the suprasegmental dimension; it can also be reflected in the segmental dimension. A growing body of research on production has shown that prosodic weight influences the amount of coarticulation between segments. Cho et al. (2017), for instance, measured the amount of coarticulatory nasality in the vowel of words such as *ban* and found less nasality when the word was carrying a pitch accent (see also Jang et al., 2018, 2023; Li et al., 2020). Another example is found in the way that the glottalization of a word-initial vowel can be modulated by prosodic weight—e.g., the vowel can be glottalized more or less in accordance with the strength of the prosodic boundary and/or stress-related prominence (see Garallek, 2013, for a review). These studies suggest that cues to prosodic structure, which determines prosodic phrasing and prominence distribution (see, e.g., Cho, 2022; Keating, 2006), are found not only in suprasegmental features such as duration and f0, but also in segmental details.

Mitterer et al. (2021b) elaborated on this possibility by asking whether listeners use such segmental details as cues to the intended prosodic structure of an utterance. To measure how listeners perceive prosodic structure, Mitterer et al. (2021b) made use of a possible mapping between prosodic and syntactic structures. This was based on the generally accepted assumption that, first, prosodic realization is influenced by the syntactic structure of an utterance through a syntax-prosody interface (e.g., Beckman, 1996; Elfner, 2018; Shattuck-Hufnagel & Turk, 1996), and second, listeners are sensitive to this relationship and exploit suprasegmental cues to syntactic structure in sentence processing (see, e.g., Kjelgaard & Speer, 1999; Schafer et al., 1996; Schafer & Jun, 2002; Snedeker & Trueswell, 2003; Steinhauer et al., 1999). Mitterer et al. (2021b) used the glottal stop in Maltese to test specifically whether segmental detail can be used to infer syntactic structure. They considered the glottal stop to be segmental because it is a phoneme in Maltese. However, it can also occur at the onset of otherwise vowel-initial words,[1] serving as a marker for prosodic structure (Mitterer et al., 2019), just as is the case in many other languages (e.g., Redi & Shattuck-Hufnagel, 2001). Mitterer et al. (2021b) asked participants to infer the intended meaning of a sequence of coordinated names, such as *Malcom u Daniel jew Gordon* (Engl., 'Malcom and Daniel or Gordon'), which could be understood as either *Malcom* [*and Daniel or Gordon*] (an early closure parsing) or [*Malcom and Daniel*] *or Gordon* (a late closure parsing). They found, in line with earlier research (e.g., Steinhauer et al., 1999), that lengthening the final syllable of the first noun, *Malcom*, in such sentences made listeners perceive a prosodic boundary after it (in line with an early closure parsing). Importantly, a glottal stop, despite the fact that it is also used as a phoneme in the language, can occur as an epenthetic one on a vowel-initial word such as /u/, the Maltese word for *and* (then realized as /u/ → [ʔu]) more likely at a larger prosodic boundary. Its presence in Mitterer et al. (2021b) indeed induced the perception of a prosodic boundary, leading participants to perceive the intended meaning as *Malcom* [*and Daniel or Gordon*]. Because the glottal stop is a phoneme in Maltese, it should be analyzed as carrying information about *what* the speaker is saying, as opposed to English, where a glottal stop is noncontrastive and thus reflects *how* a speaker is producing a word. The result, as Mitterer et al. (2021b) argued, shows that listeners use segmental information to make choices about the prosodic structure, which is, in this case, conditioned by the syntactic structure.[2]

In this paper, we continue to investigate the use of segmental detail as a cue to prosodic structure by examining another case of the role of segmental detail with a segmental reduction in the German auxiliary *haben* /habən/ (Engl. 'to have'). This study not only adds to a relatively small body of research on the role of segmental detail in processing prosodic structural information in relation to other higher-order linguistic structures, but it also complements Mitterer et al.'s (2021b) results, which are open to alternative interpretations. One of the auxiliary assumptions of Mitterer et al. (2021b), linking the data to theoretical claims, was that the glottal stop is segmental, given its phonemic status. Nevertheless, we cannot entirely rule out the possibility that the glottal stop might not be taken to be purely segmental. In fact, parallel investigations into the representation of the glottal stop in Maltese (Mitterer et al., 2021a)[3] brought up the possibility that, in line with some linguistic accounts (Kehrein & Golston, 2004), the glottal stop might still be prosodic, even when it is considered a phoneme. This inter-

[1] Note that the distinction between vowel-initial and glottal-stop initial words is one of underlying phonological representation, and not just orthography, because different phonological processes associated with vowel-initial versus glottal-stop-initial words in Maltese (see Mitterer et al., 2019).

[2] A caveat in this context pertains to the treatment of the glottal stop as a segmental entity. Glottalization, which is often used interchangeably with laryngealization, may be regarded as a form of laryngeal modification as a suprasegmental feature, as often found, for example, with low pitch in falling intonation or a signal for a prosodic boundary (e.g., Lehiste, 1970). However, Mitterer et al. (2021b) view that a fully realized glottal stop should be classified as part of the segmental inventory in Maltese, where it is used as a phoneme, as well as functions as a prosodic marker. Given that a full epenthetic glottal stop is considered to be inserted before a vowel, similarly to a phoneme, it is regarded as a segmental entity in this context.

[3] Note that our work on the epenthetic glottal stop preceded our work on the lexical representation of the glottal stop, so we were not aware of the results in Mitterer et al. (2021a) when interpreting the results of Mitterer et al. (2021b).

pretation is based on the result that the glottal stop in Maltese, despite being a phoneme, does not strongly constrain lexical access as the oral stop /t/ does (Mitterer et al., 2021a, Experiment 3). This possibility calls for further studies to buttress the view that segmental detail, which is generally taken to provide the *what* component of a linguistic message, can serve as a cue to prosodic structure, thus contributing to the *how* component.

The present study also expands Mitterer et al. (2021b) in terms of the interface between prosodic structure and higher-order linguistic structures. Whereas Mitterer et al. (2021b) made use of the relationship between prosodic structure and syntactic structure, in this study, we exploit the relationship between prosodic structure and information structure. That is, listeners' prosodic processing is measured in relation to prosodic structure as a cue for the information structure rather than the syntactic structure. This auxiliary assumption is supported by ample evidence. Speakers create a particular prosodic structure of prominence to hyperarticulate (or accentuate) linguistic units that emphasize new or contrastive focus information, in accordance with the information structure (e.g., Cangemi & Baumann, 2020, and the references therein; Krahmer & Swerts, 2001), and listeners interpret those units as such (e.g., Dahan et al., 2002; Ito & Speer, 2008; Weber et al., 2006). In a visual-world task with eye-tracking, Weber et al. (2006), for example, showed that listeners interpret a pitch accent on an adjective in a noun phrase as meaning that color is crucial in determining the intended referent. A pitch accent on 'red' is thus taken as an indication that the intended referent of that noun phrase has a competitor on the screen that only differs in color (i.e., when there is a red and a green ball), and participants look to that object even if another red object is available that does not have a differently colored competitor on the screen.

Methodologically, we base our investigations on a mouse-tracking paradigm that was used by Roettger and Franke (2019), but we adapted the paradigm to a remotely accessible online format. Roettger and Franke (2019) also tested the role of prominence in predicting an upcoming referent. As noted by Cangemi and Baumann (2020), prominence may refer not only to the concept of 'standing out,' where a stressed syllable stands out from other syllables within a word through cues such as loudness and duration, but also to 'standing out' of constituents (whether they are syllables or words) from other referents within a phrase. This latter distinction is typically achieved through the choice and placement of pitch accents, which is often conditioned by focus. Focus, in this context, is an information structural notion that designates a specific part of an utterance assumed to convey important information in a given context (see Mücke & Grice, 2014, and references therein). For example, in a Wh-question like 'What did they collect?', an appropriate answer, such as 'They collected a violin,' should have 'a violin' corresponding to 'What' as the most informative part of the utterance. This type of focus is often referred to as 'narrow' focus because it narrows the required information down to a specific constituent, in this case, 'a violin.'

However, narrow focus can also be achieved in a context where the focused constituent contrasts with what has been previously mentioned by the other interlocutor. For instance, when someone answering a question like 'Did they collect a kettle?' believes it was not a kettle but a violin, they can respond with 'No, they collected a VIOLIN.' In such a case, the contrasting constituent 'violin' is referred to as 'contrastive' focus. In contrast to these focus types placed on a 'narrow' domain, an utterance as a whole may be said to receive 'broad' focus if the entire utterance conveys new information. Such a context arises when an utterance like 'The aliens collected a violin' is in response to a question like 'What happened?'.

Yet another special type of focus relevant to the present study is called 'verum' focus in German. It generally refers to a semantic effect of information structure that emphasizes the expression of the truth of a proposition (Lohnstein, 2016). Consider the following short dialogue. In German, the auxiliary *haben* would receive a pitch accent (i.e., v*erum* focus) in (1b) if the speaker wanted to indicate an affirmative response to the yes/no question in (1a) (Turco et al., 2014):

(1a) *Haben die Aliens die Geige eingesammelt?*
have[3PL] the[F] aliens the[F] violin collected?
'Did the aliens collect the violin?'
(1b) *Die Aliens haben dann die Geige eingesammelt.*
the[F] aliens have[3PL] then the[F] violin collected.
'Then, the aliens did collect the violin.'

In general, as discussed in Mücke & Grice (2014), the different types of focus may carry varying levels of prominence, arranged in increasing order as 'broad' focus, 'narrow' focus, and 'contrastive' focus. As for 'verum' focus, it is not clear exactly what level of prominence it may carry. However, a recent study found that German listeners find pitch accents with larger pitch movements (L + H*, L*+H) more appropriate for verum contrasts than an H*, even more so than for contrastive focus (Röhr et al., 2023), which indicates that it may be similar to or even stronger than contrastive focus.

Roettger and Franke (2019) particularly tested the listener's use of the German *verum* focus in predicting an upcoming referent. They showed that listeners indeed use the *verum focus,* expressed through prominence on the auxiliary verb *haben*, in sentences to predict the remainder of the sentence in a mouse-tracking paradigm. Participants were presented with a screen containing two objects, for instance, a violin and a pear (similar to the set-up in Fig. 2), one of which was mentioned in the question. Then a mouse cursor, represented by a yellow circle, appeared, and when participants clicked on the circle, the answer was heard. Participants had been instructed to listen to the dialogue and move the yellow circle to the object that would be collected according to the answer sentence. The results showed that when the auxiliary *haben* received a verum focus (prosodically marked with a pitch accent), participants moved the circle more quickly to the object that had been mentioned in the question sentence than when it did not receive a verum focus.[4]

Roettger and Franke (2019) implemented the *verum* focus by using suprasegmental features such as f0 contour and

[4] Roettger and Franke (2019) also used a condition in which the speaker used prosody in a way that is atypical for German (i.e., *verum* focus when the object to be collected was the other one). This condition can be ignored for the present purposes.

amplitude of the auxiliary verb. In the present study, we test the extent to which the segmental detail of the auxiliary verb, which also varies with prominence, can be exploited by the listener to anticipate the following referent in a dialogue with a given information structure. The plural form of *haben* is useful for this purpose because its segmental realization shows variable phonetic forms that can be described as having more or less coarticulation (or segmental reduction) of the individual segments, as shown by an analysis of the pronunciation of *haben* in the Kiel Corpus (IPDS, 1994), a corpus of spoken German. The corpus contains 210 utterances of *haben*, of which 55 occur in read speech and 155 occur in spontaneous speech, with three major phonetic forms: [habən], [habm], and [ham]. The full form is found only twice in read speech (constituting 3.6% of the items) and never in spontaneous speech. Given the low frequency of occurrence of the full form, it is clear that the full form is marked, drawing the listeners' attention more than other reduced forms. This, in turn, might lead to the inference that the speaker is placing prosodic weight (in this case, prominence) on the auxiliary, which might, in turn, suggest that the auxiliary verb is carrying verum focus. Furthermore, in read speech, the form without a schwa and nasal place assimilation [habm] (/n/ is realized as [m] after /b/) is the most common (accounting for 87.2% of the tokens), with the strongly coarticulated (thus reduced) form [ham], making up 9.1% of the tokens. In spontaneous speech, there is a preference for this strongly coarticulated form (accounting for 58.7% of the tokens, with 41.3% for [habm]).

Because our experiments were run in a remotely accessible online format, we had to make some modifications from the original experiment of Roettger and Franke (2019) (see Experiment 1 for the details).[5] Therefore, in our first experiment (Experiment 1), we tested whether our online version of the mouse-tracking paradigm, with the voice of a new German speaker, would work and successfully replicate their findings. For this purpose, we manipulated prosodic (suprasegmental) parameters, f0, and amplitude to create different prosodic versions of *haben* in the answer sentences, as was the case in Roettger and Franke (2019). In Experiment 2, we investigated whether the segmental detail of phonetic forms of *haben* that differed in the amount of coarticulation (segmental reduction) would lead participants to predict the upcoming referent in line with the information structure embedded in the prosodic-structurally conditioned segmental detail. In Experiment 3, we ran a control experiment to further evaluate whether the results obtained in Experiments 1 and 2 simply reflected task-related learning effects or could be reliably attributed to the participants' computation of a prosodic structure in relation to the information structure given.

2. Experiment 1

In Experiment 1, as just mentioned above, we tested whether a remotely accessible online version of the mouse-tracking paradigm would show results comparable to those in a lab-based study (Roettger & Franke, 2019)—i.e., listeners use prominence-related suprasegmental cues to prosodic structure (indicating a verum focus) in predicting an upcoming referent. One issue in mouse-tracking experiments is how to keep participants motivated to move the mouse in anticipation of the imperative stimulus, that is, the critical word of the sentence that gives away the target to be clicked. In lab-based mouse-tracking experiments, anticipatory mouse movements are often encouraged by slowing down the maximum speed of the cursor, as was done in Roettger & Franke (2019). That leads participants to start moving the cursor before they hear the imperative stimulus because otherwise, they feel it would take an unnecessarily long time to reach their intended target (and therefore it would take a longer time to complete the task). But we could not manipulate the speed of the cursor in the online setting, so instead, we gamified the task, as is often recommended for online experiments (Hartshorne et al., 2019), to resemble the classic arcade game Galaga (see Section 2.1.3 for details).

Following Roettger and Franke (2019), we employed three conditions of question–answer sequences. In the baseline condition, the question was 'Was ist passiert?' (Engl. 'What happened?') which did not contain any noun phrase that may be related to the target word. Consequently, the answer featured a broad focus context characterized by a non-committal intonation pattern with slight declination. This served as a baseline to measure how quickly listeners could move the mouse to the target upon hearing the target word in the answer sentence. We anticipate faster responses in the so-called *verum* and *contrast* conditions. In both of these conditions, the question asks whether a specific object has been collected, as exemplified in (1a), and the answer confirms it. In the *verum* condition, there is a pitch accent on the auxiliary, while in the *contrast* condition, the pitch accent is on the object name. If our online version is capable of capturing the influence of prosody on sentence processing, we would expect to observe quicker responses in the verum condition compared to the other two conditions and a smaller advantage of the contrast condition over the baseline. In other words, when listeners hear the pitch accent on the auxiliary verb, they would interpret the response as being affirmative to the question, making it more likely for the target word to be the same as the noun mentioned in the question sentence. On the other hand, when listeners hear no pitch accent on the auxiliary verb, they would interpret the response as being contrastive with the question, making it more likely for the target to be different from the one mentioned in the question sentence. Listeners' interpretations of these experimental conditions should be reflected in performance differences relative to the baseline condition, where the intonation does not convey specific information about the givenness of the target. Furthermore, it's important to note that these effects become more pronounced over the course of the experiment, as demonstrated in Roettger and Franke (2019).

2.1. Method

2.1.1. Participants

Twenty participants were recruited from the Prolific platform on the conditions that their native language must be German and that their age must be in the range between 18 and 40.

[5] Lab-based mouse tracking allows control of the mouse's position and maximum speed, which is not possible in an online setting. This is because JavaScript, fortunately, does not allow any control of the cursor. One would not want a website to move the mouse cursor over an advertisement and then reduce the maximum speed.

In the sign-up information, participants were informed that the experiment made use of mouse-movements in a simple game-like design. The actual ages of participants ranged from 20 to 36, with a median of 24.5 years. Twelve of the participants were male and eight were female. Roettger and Franke (2019) had about 30 participants in their experiment in the "reliable" prosody condition that we here aimed to adapt to an online setting. It might be considered unusual to have a smaller sample size than the original study, but it is important to note that the primary purpose of Experiment 1 is not to replicate Roettger and Franke (2019) but to test whether their paradigm could be adequately adapted to a remotely accessible online format. Furthermore, given their large effect sizes (more than 100 ms differences between conditions), a smaller sample size of 20 participants were deemed sufficient to reveal an effect that would indicate successful adaptation.

2.1.2. Material

We used the same visual materials for the 12 objects as Roettger and Franke (2019), which were made available on OSF (Roettger & Franke, 2022). For the auditory material, we recorded a female speaker uttering the general question (*Was ist passiert?*) used for the baseline condition and the twelve questions used for the *verum* and *contrast* conditions, asking about whether one of the twelve objects had been collected (e.g., *Haben die Aliens die Geige eingesammelt?* Engl., 'Have the aliens collected the violin?'). A male speaker recorded all the answers (i.e., *Die Aliens haben dann die* OBJECT *eingesammelt?* Engl., The aliens have then collected the OBJECT.') with the three different intonations multiple times for each of the twelve objects. One of the most neutral sounding utterances, as confirmed by the first author, was selected as the base to generate baseline, contrast, and verum versions using the PSOLA algorithm in Praat (Boersma, 2001). Utterances were selected to have a starting pitch around 150 Hz that fell to around 110 Hz at the end of the utterance, showing some modest degree of f0 declination.

For the baseline version, all pitch points for the auxiliary *haben* and the object name were removed, and the pitch curve was linearly interpolated for those words based on the preceding and following pitch values. The resynthesis of this file was used as the baseline. This was done so that not only the prosodically marked condition but also the baseline condition was created with a PSOLA resynthesis, ensuring that the conditions differed only in their prosody. For the *verum* condition, an L + H* type of contour was implemented on the auxiliary verb. To achieve this, the first pitch (f0) point during the auxiliary verb was lowered to 90% of the original value obtained with the f0 interpolation, and at the amplitude maximum, the pitch was set to 210 Hz, with two points at 90% of this pitch maximum set 20 ms before and after the amplitude maximum to generate an f0 plateau. For the *contrast* condition, the pitch (f0) contour for the object name was manipulated in a similar way around the maximum amplitude of the word but with a maximum of 230 Hz. These f0 values were based on the typical values observed in natural utterances. Fig. 1 shows the resulting stimuli with their respective pitch contours for one of the targets (all stimuli are available via OSF). After PSOLA manipulation, the files were saved as WAV files and then converted to MP3 using Audacity software using an *Extreme Preset* for minimal compression. (Note that all files are transferred to the participants' computer at the start of the session, hence a large file size does not pose a timing issue.).

2.1.3. Procedure

The experiment started with an instruction screen that explained the task with an example display and written text. It explained that the participants would hear a dialogue between two speakers who are observing an alien spaceship and explained the procedure.

Fig. 2 provides an overview of how individual trials were implemented. The display starts with two objects located right and left on the top of the screen and a spaceship at the bottom in the middle. In lab-based settings, the mouse position is usually forced to a starting position at the start of a trial, but this is not possible in an online setting. Therefore, participants have to click on the spaceship to start the trial. In this way, we control the starting position of the mouse. With this click, three things happen: First, a dialogue similar to the one exemplified in (1) begins to play. Second, in synchrony with the onset of the answer sentence, the two objects start to fall down, slightly wiggling left or right. Third, the position of the spaceship is now controlled by the mouse's horizontal movement. That is, participants can move the spaceship to the left and right. Participants received instructions on the procedure through written instructions displayed on the screen. Their task was to move the spaceship swiftly to position it beneath the falling object that corresponded to the correct answer. As illustrated in the lower panel of Fig. 2, participants were then required to click a mouse button to activate a 'tractor beam' (a device commonly found in science fiction) to intercept and draw the object towards the alien spaceship. In line with the gamification aspect, we provided audio feedback to indicate the accuracy of their choice: a 'bling'-like sound for correct selections and a buzzer sound for incorrect ones. A choice was also deemed incorrect if the 'tractor beam' failed to capture the falling object, even when the spaceship was positioned on the side of the screen where the target was located. This accuracy requirement, coupled with the unpredictable wiggling movements of the falling objects, was intended to encourage participants to make quick decisions about which object to track.

The experiments were developed in PsychoPy (Peirce et al., 2019) and run on the Pavlovia platform. The experiment started with a welcome screen that lasted for 100 frames to estimate the frame rate of the participant's monitor. To account for different screen sizes, the layout was arranged in PsychoPy using normalized units that ranged from −1 to 1.

After the instruction, participants completed twelve example trials in which each object was the target once to familiarize them with the objects and their names. After that, twelve blocks of twelve trials, for a total of 144 trials, were completed. Each trial started with two objects, sized to fill roughly 10% of the screen width and height, initially positioned at the top of the screen with their midpoint at 92% of the screen height (Fig. 2). The objects, timed with the acoustic onset of the answer sentence, fell down the screen at a rate of 15% of the screen size per second. If the participant failed to respond on a given trial, the objects continued to fall until they reached 40% of the screen height, at which point they would nearly reach the spaceship. This means that the maximal falling time

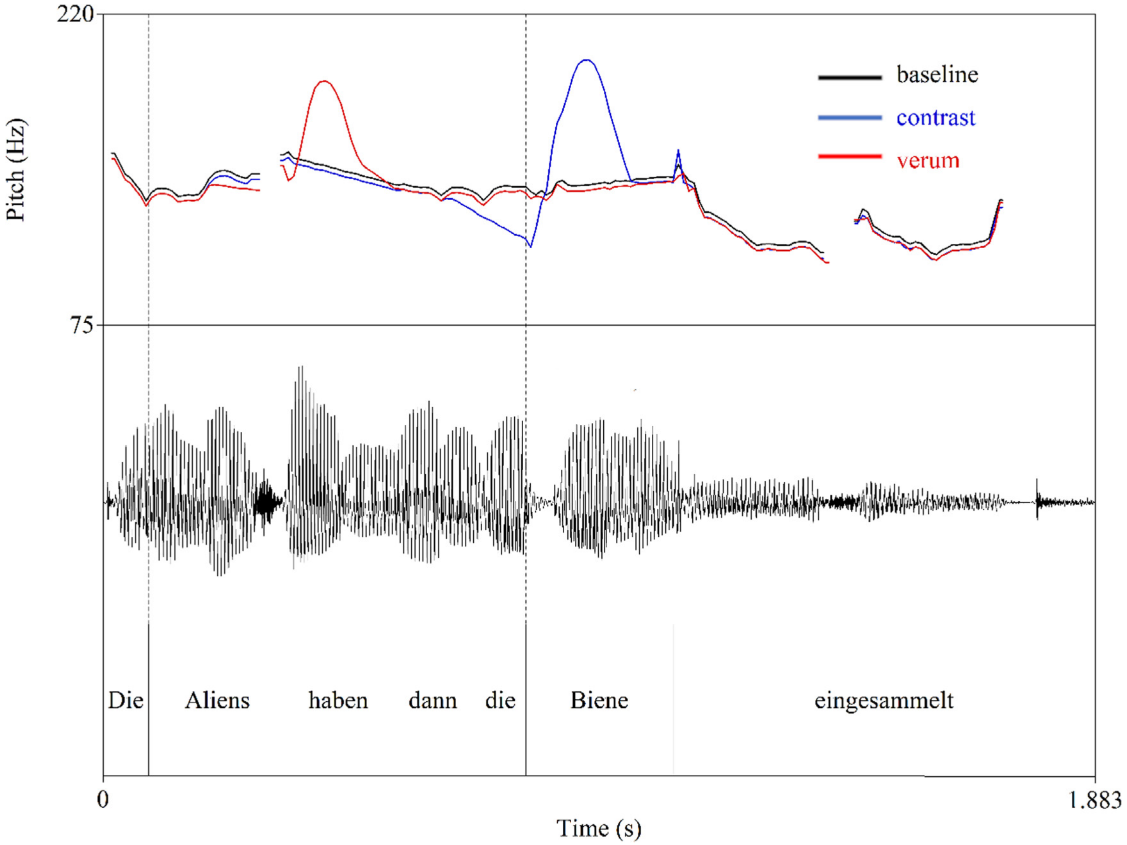


Fig. 1. An example of a stimulus triplet based on the sentence 'Then, the aliens did collect the bee' (see example 1b for a gloss). The middle panel shows the waveform, and the upper panel displays the pitch contours, with the baseline adjusted 2 Hz upwards from its actual position for visibility. In the ***verum*** condition, a pitch accent is placed on the auxiliary 'haben,'; in the ***contrast*** condition, a pitch accent is placed on the object name 'Biene' (Engl., 'bee'); and in the **baseline** condition, all pitch points for the auxiliary 'haben' and the object name were removed, resulting in a linear interpolation of the pitch curve for those words based on the preceding and following pitch values.

was just less than 4 s (to cover 52% of the range at a rate of 15% per second). The objects were not only moving down but also wiggling to the left and right. The starting position of the two objects was at 20% and 80% of the screen width. The direction of the wiggling changed randomly, with a 2.5% chance of a direction change with every frame. But the direction would also change if an object would otherwise fall out of a corridor of 20% of the screen width centered around the starting position (as indicated by the grey rectangles in Fig. 2). The trial ended when the participant clicked a mouse button and thus activated a tractor beam to capture the object for collection, as shown in Fig. 2b (lower panel).

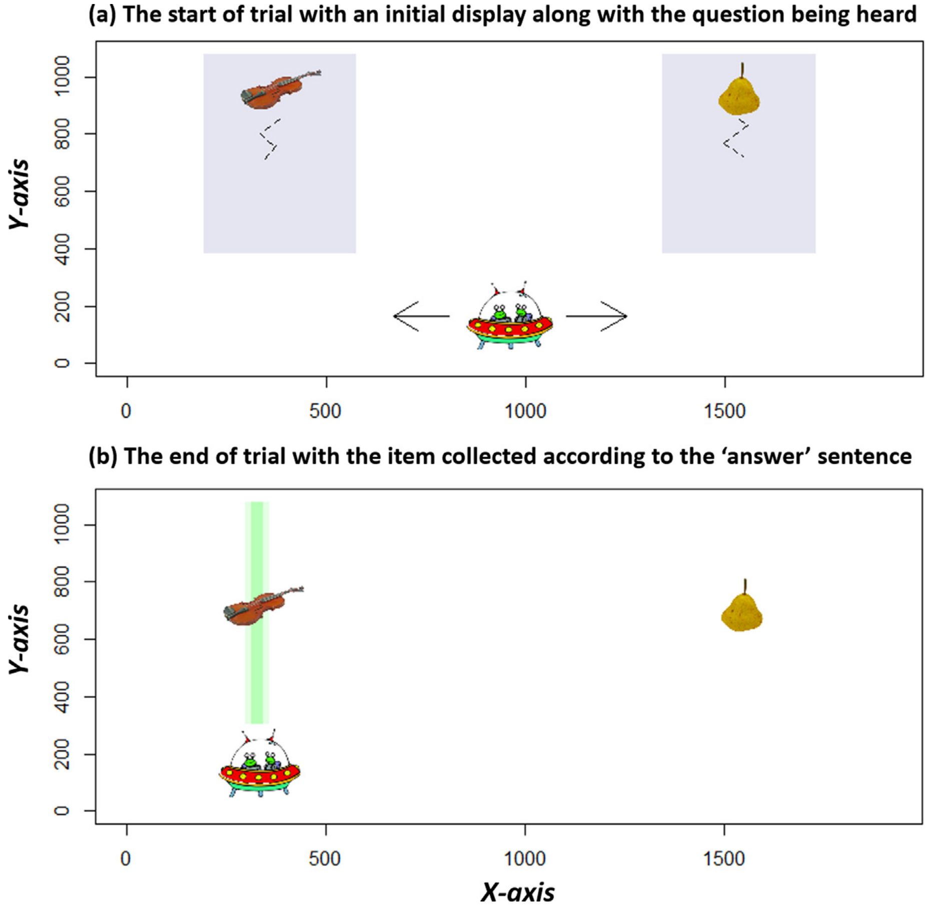
Each block contained four trials for each of the three prosodic conditions—i.e., a verum, a contrast, and a baseline condition. The assignment of objects to conditions was rotated across the blocks so that each object occurred four times in each condition. Randomization was done offline, that is, we prepared 40 different trial sequences and one of those was chosen at random for a given participant. The randomization did not allow target-competitor pairs to share the same onset consonant (e.g., *Hose – Henne, Birne – Biene*). After every thirty trials, participants were told their average accuracy and reaction time for correct trials to encourage them to stay engaged in the task.

2.1.4. Data preprocessing and analysis

For the mouse tracking, the data were preprocessed as described below (all files available on OSF: https://osf.io/v8yqn/). All pre-processing and statistical analysis was done in R v4.2.1 (R Development Core Team, 2022).

The output contains arrays about the mouse position and the times at which those mouse positions were observed. To normalize those observations and average them across trials, we generated a timeline from 500 ms before the target word onset to 2000 ms after the target word onset, with steps of 16.6 ms (=60 Hz) for each trial. For each of those time points, the nearest observed mouse position was used. Note that this also means that if a participant reacted before 2 s had elapsed, the timeline was padded with the last observed mouse position. The mouse positions were coded in normalized screen coordinates and were thus in the interval [-1,1]. To facilitate the analysis, we inverted the data for trials in which the target was on the left, so that all values above zero indicate that the mouse position is on the target side.

For the statistical analysis, we used two dependent variables, as common in eye-tracking research, following the model often used in eye-tracking research (Steffman & Sundara, 2023). First of all, we measure reaction time, defined as the duration from the acoustic onset of the target to the


(a) The start of trial with an initial display along with the question being heard

(b) The end of trial with the item collected according to the 'answer' sentence

Fig. 2. Task display on a 1920x1080 screen. The upper panel (a) shows the initial display indicating the corridor (as marked by the semi-transparent rectangles which were not visible during the experiment) in which the objects were falling down and wiggling left and right as indicated by the dashed lines. As the arrows indicate, the spaceship could only be moved in the horizontal direction. The lower panel (b) shows the display as the participants move the spaceship under one object and activate the tractor beam with a mouse click.

moment of clicking the mouse button to shoot the object. This indicates the time at which participants clicked the mouse button (to activate the 'tractor beam') in order to collect the object. Reaction times provide a clear indication of how difficult participants found the task in different conditions. For the reaction (mouse click) times, we analyzed only trials in which participants used the mouse button to activate the tractor beam. (Note that reaction times were measured relative to the acoustic onset of the target word. This approach was chosen because it provides a consistent point in time for measurement across all conditions, including the baseline condition where the auxiliary does not reveal the target.) Secondly, we use a measure that more purely reflects early processing based on the mouse-tracking data. We use the same measure as Roettger and Franke (2019), who called this TTT (turn towards target) and we call this measure decision time, to contrast with reaction time. Both times are measured relative to the onset of the target word in the answer. This "decision time" was found to be the time during which the participant started moving the mouse cursor toward the target, entered the target corridor, and *continued* to stay inside the corridor before collecting the target (Fig. 2). Note that the decision time cannot be determined if participants guess early and move to the target before the start of the time window. Those cases were therefore not used in this analysis (see Section 2.2.2 for more on this).

All analyses made use of linear mixed-effects models using the package lmerTest (Kuznetsova et al., 2015), with participant and item as random effects. Modeling started with a maximal random effect structure, including all possible random slopes and their correlations. When the maximal models did not converge, they were iteratively simplified until convergence was reached. Effects were considered significant if $p < .05$. For the three-level condition variable, the first contrast compared the baseline level with the two potentially informative levels (*verum* vs. *contrast,* with the latter mapped to 1/3 and the baseline condition mapped to −2/3). The second contrast compared the latter two conditions with each other, with the *verum* condition mapped to −0.5 and the contrast condition to 0.5. Based on the findings of Roettger and Franke (2019), we should, therefore, expect negative regression weights when the more informative conditions lead to faster reaction and decision times. The Block predictor ranged from zero to eleven, so the intercept values reflect the estimated differences at the start of the experiment. Negative regression weights for the interaction of Block with the condition predictors would indicate that the advantages further increased during the experiment, while positive regression weights would suggest that the advantages decreased during the experiment (given negative regression weights for the main effect of the condition variable).

2.2. Results

2.2.1. Reaction times

Recall that reaction time is defined as the duration from the acoustic onset of the target to the moment of clicking the mouse button to collect the object. We first tested whether participants completed a sufficient number of trials with button presses. Two participants did not: one never used the mouse button, and one used it in only 33% of the trials. Those two participants were excluded from the analyses. In the remaining 2478 trials, all other participants used the mouse button on more than 90% of the trials, and only 14 trials were rejected for missing responses. 14 participants (of the remaining 18) always used the mouse button.

For the remaining trials, we used an intercept-only linear mixed effect model with participant and item as random factors to find the trials in which the reaction time had a normalized residual larger than an absolute three, leading to the rejection of an additional 38 trials (1.3%). Fig. 3 shows the mean reaction times for the three conditions as they evolved across the twelve blocks.

The results were then analyzed with a linear mixed-effects model with Block, Condition, and their interaction as predictors (see the section Data preprocessing and Analysis for the predictor coding). The maximal model that converged excluded correlations between random slopes and contained random slopes for both contrasts over participants but, over items, only the random slope for the contrast *Baseline vs. Informative* remained. Random slopes for Block and its interactions had to be removed. The results (Table 1) show that at the start of the experiment, the combined *verum* and *contrast* conditions had an advantage over the baseline condition, reflected in a negative regression weight for the first contrast (Baseline vs. Informative predictor). There also was an advantage of the *verum* condition over the *contrast* condition (whichInformative predictor). Both effects grew larger over the course of the study, but significantly so only for the first contrast.

Table 1
Results of the linear mixed-effects model for the reaction times of Experiment 1. 'Informative' refers to both the *verum* and *contrast* conditions combined where an intonation contour for a pitch accent is realized on the auxiliary verb and the following target noun, respectively, whereas the intonation contour was removed from these two words in the baseline condition.

	B (SE)	t (df)	p
Intercept	1074 (51)	20.997 (19)	<0.001
Block	−11 (1)	−7.972 (2365)	<0.001
Condition			
Baseline vs.Informative	−89 (27)	−3.305 (42)	0.002
whichInformative	−71 (30)	−2.347 (44)	0.023
Condition x Block			
Block:(Baseline vs. Informative)	−6 (3)	−2.142 (2382)	0.032
Block:(whichInformative)	−1 (3)	−0.377 (2380)	0.707

Note: The syntax of the final model was (1000*RT_targetOnset) ~ block*(baseLineVsInformative + whichInformative) + (1 + baselineVsInformative + whichInformative ||participant) + (1 + baselineVsInformative||item).

2.2.2. Mouse-tracking data and decision times

After preprocessing, we checked whether each participant had a detectable decision point in at least 50% of the trials—i.e., whether the participant moved the mouse towards the object in the 500 ms – 2000 ms window, which was used to calculate a decision time, for at least half the trials. (Recall that the decision time was determined by the time during which the participant began moving the mouse cursor toward the target, entered the target corridor and *continued* to stay inside the corridor before collecting the target.) This criterion differed from the one used to exclude the two participants who did not provide a sufficient number of mouse clicks overall. For this analysis, all participants could be retained.

Fig. 4 shows the raw mouse positions relative to the target-word onset. The left panel shows all trials. Here, note that in the pre-target time window (−500 to 0) in Fig. 4a, the x-value was higher than 0 in the *verum condition*, indicating that the mouse was positioned toward the target side, whereas the x-value was lower than 0 in the *contrast* condition, indicating that the mouse was positioned toward the competitor side,

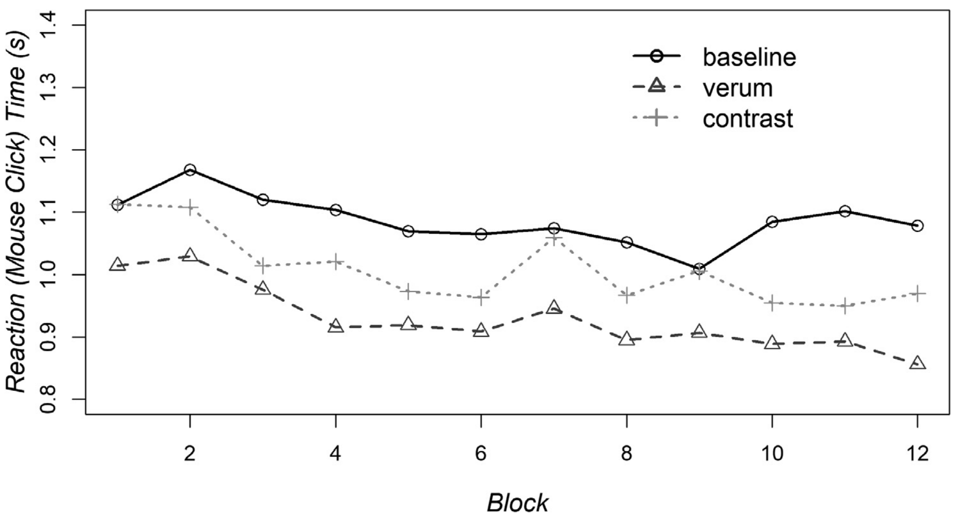


Fig. 3. Mean reaction (mouse click) times (to collect an object) relative to the acoustic onset of the target word in the three conditions (baseline, verum, contrast) of the twelve blocks in Experiment 1.

although it was eventually moved toward the target. In other words, there was a bias toward the target in the *verum* condition but a bias toward the competitor in the *contrast* condition. Recall that in the question sentence (see (1)), the target was mentioned in the *verum* condition, and the competitor was mentioned in the *contrast* condition. Thus, it appears that participants were biased to go toward the object mentioned in the question. Although that could be interpreted as an effect of the information structure, we attempted to correct for that bias by removing 185 trials (6.7%) of early guesses, in which the mouse position was already in the target corridor during the pre-target time-window (from −500 to 0). Fig. 4b (right panel) shows the data after we removed those trials. This figure shows that participants were able to use the presence or absence of a pitch accent on the auxiliary verb to move toward the target more quickly than in the baseline condition, even after rejecting the early guesses.

To statistically analyze these data, we calculated the decision time for each trial (see Section 2.1.4). Fig. 5 shows how the decision times for the three conditions developed across the twelve blocks, and Table 2 summarizes the results of the statistical analysis. The maximal model that converged

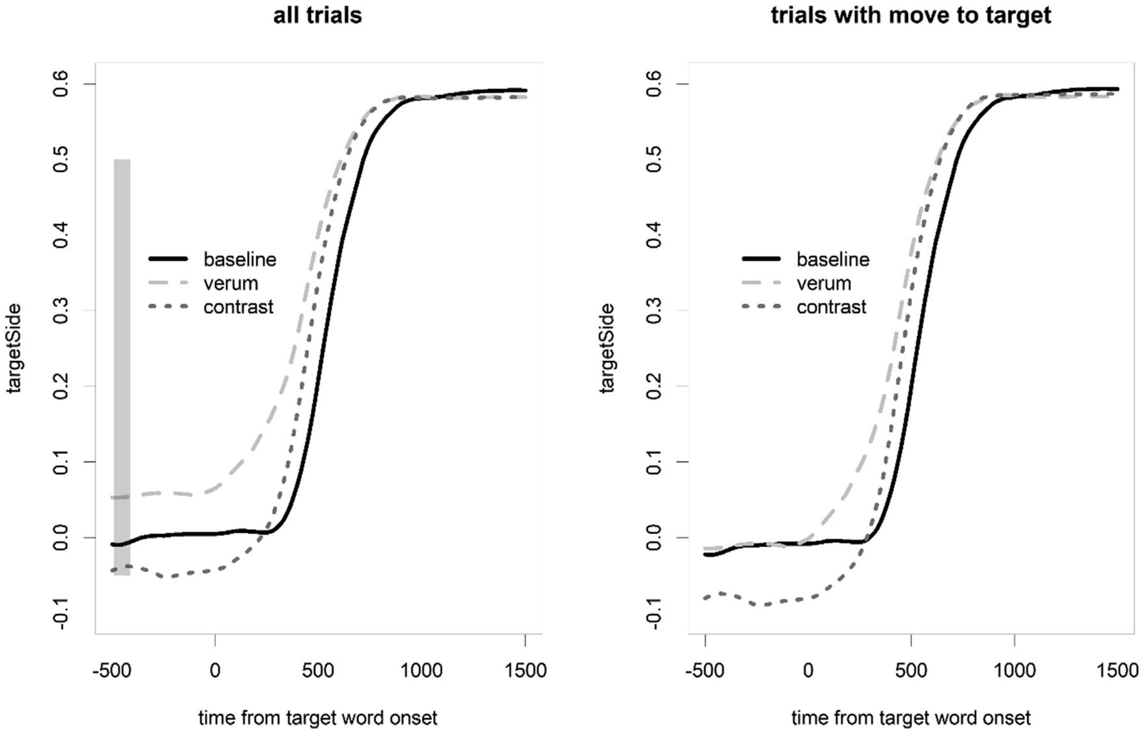


Fig. 4. Mouse positions in Experiment 1 in the time window from −500 ms to 1500 ms relative to the target word onset. The left panel (a) shows all data and the right panel (b) only trials with a move towards the target after removing data (185 trials, 6.7%) in which the mouse was on the target side during the whole pre-target window. The gray area on the left panel (a) indicates the range of endpoint of the auxiliary *haben* (containing potentially informative prosodic cues) in the answer sentences relative to the onset of the target word.

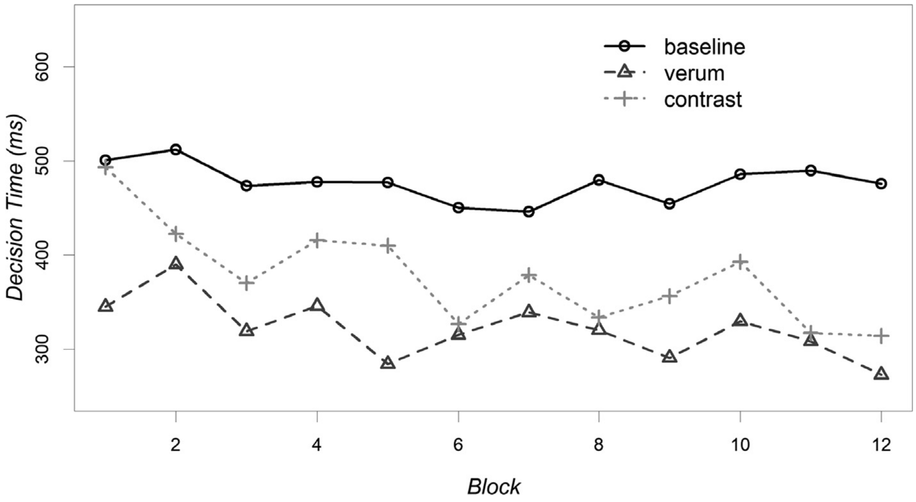


Fig. 5. Decision times in the mouse tracking data in Experiment 1. The decision times were calculated from the time at which the participant started moving the mouse cursor towards the target to the time at which the mouse entered the region (corridor) of the target.

Table 2
Results of the linear mixed-effects model for the estimated decision times of Experiment 1.

Predictor	*B (SE)*	*t (df)*	*p*
Intercept	424 (37.4)	11.343 (22)	<0.001
Block	−6 (1.2)	−5.062 (2523)	<0.001
Condition			
Baseline vs. Informative	−98 (24.1)	−4.071 (43)	<0.001
whichInformative	−85 (22.1)	−3.829 (90)	<0.001
Condition x Block			
Block:(Baseline vs. Informative)	−7 (2.5)	−2.662 (2531)	0.008
Block:(whichInformative)	5 (3.0)	1.605 (2539)	0.108

Note: The syntax of the final model was: decision ∼ block*(baselineVsInformative + whichInformative) + (1 + baselineVsInformative + whichInformative||participant) + (1| item).

excluded correlations between random slopes and contained random slopes for both contrasts over participants but only a random intercept over items. Random slopes for Block and its interactions had to be removed.

The patterns in Fig. 5 and the statistical results regarding decision times are similar to those observed for reaction times (Fig. 3 and Table 1), indicating that the two conditions (*verum* and *contrast*) with informative pitch accents exhibit an advantage over the baseline condition. Furthermore, the *verum* condition leads to faster decision times compared to the contrast condition (refer to the *whichInformative contrast* in Table 2). In other words, when the pitch accent was on the auxiliary verb *haben* (the *verum* focus condition, marked by a dashed gray line in Fig. 5), participants moved the mouse more quickly to the target mentioned in the question. Conversely, in the *contrast* condition (marked by a dotted line in Fig. 5), when the pitch accent was on the target noun itself, participants moved the mouse more quickly to the target not mentioned in the question (thus contrasting with the one that had been mentioned), compared to the baseline condition.

It is also worth noting that we found a significant interaction between Condition and Block for the contrast between the baseline condition and the two conditions with an informative f0 contour. This means that the advantage of the two informative conditions over the baseline condition increased over the course of the experiment (see the regression weight for 'Block:[Baseline vs. Informative]' in Table 2). Although Fig. 5 suggests that this growth was more pronounced in the contrast condition, the interaction term for the difference between the two conditions with informative intonation and Block did not reach statistical significance.

2.2.3. Discussion

The results of Experiment 1 are generally in line with those that Roettger and Franke (2019) obtained in a laboratory setting. That is, we successfully adapted the mouse tracking paradigm to a remotely accessed format. Crucially, participants were found to use the presence of a pitch accent on the auxiliary verb in reference to the information structure, predicting that the target object would be the one previously mentioned in the question. Interestingly, the absence of such an accent appears to be informative because it led participants to choose the not-previously-mentioned target more quickly in the *contrast* condition than in the baseline condition, even though that advantage is numerically much smaller than the one for the verum-focus condition. As predicted, when there was no clearly realized pitch accent on the preceding auxiliary verb, listeners appeared to interpret the response negatively, that is, contrasting with the question. This made it more likely for the subsequent target noun to be different from the one mentioned in the question sentence. Because there are only two options, and this rules out the object mentioned in the question, this increases the predictability of the target. Moreover, since our randomization procedure ensured that the target and competitor were easily distinguishable by the first consonant (i.e., a target like 'Biene' was never accompanied by a similar-sounding word like 'Birne'), the pitch accent on the target word contributed little to finding the target (Allopenna et al., 1998). The advantage in the contrast condition, however, should also be interpreted with caution because it could be due at least in part to learning during the task—i.e., the advantage of the *contrast* condition over the baseline condition got stronger over the course of the experiment. However, the advantage of the *verum* condition did not increase over the course of the experiment, indicating that the pitch accent effect on the auxiliary *haben* is independent from a task-specific learning effect.

Given that we established a solid pitch accent effect, we next moved on to Experiment 2 to test whether a similar effect could be observed with a prosodic-structurally conditioned segmental detail that, in this case, related to how strongly the segments of the German auxiliary *haben* were reduced by coarticulation (weak reduction: [habən], medium reduction: [habm], strong reduction: [ham]).

3. Experiment 2

In Experiment 2, we investigated whether the segmental detail of the auxiliary *haben* would lead participants to predict the upcoming referent in accordance with the information structure embedded in a prosodic-structurally conditioned segmental detail—i.e., the full form [habən] in the *verum* focus condition, the strongly reduced (coarticulated) form [ham] in the *contrast* condition (in which the following noun for the target is contrastive to the one previously mentioned in the question); and the weakly reduced form [habm] in the baseline condition (most frequent, cf., the Kiel Corpus (IPDS, 1994)). Especially the low frequency of the full form might indicate to listeners that the speaker is putting prosodic weight on the *verum* focus of the auxiliary even in the absence of an f0 cue.

It is worth noting that we used the term 'segmental detail' of the auxiliary verb as a cover term to encompass its three different forms, which involve segmental deletion and coarticulatory differences. The phonetic realizations of these forms may also exhibit differences in suprasegmental dimensions, particularly in duration. For instance, the reduced form is likely to be shorter than the full form, not only due to segmental reduction but also because of the difference in the number of syllables (as observed in our materials; see Fig. 6). Similarly, the strongly reduced form '[ham]' may also be shorter than the weakly reduced form '[habm]' due to distinct syllable structures. In other words, our manipulation of 'segmental' detail also results in a change in duration, which is a suprasegmental feature. Thus, we assume that any observed effects associated with the different forms may result from a combined effect of their segmental and suprasegmental characteristics.

Nevertheless, it is important to reiterate that this segmental variation, which may be accompanied by some durational modification, is conditioned by prosodic structure because it reflects the prominence component of the prosodic structure. The full form reflects hyperarticulation and prosodic strengthening of segmental realization, which usually accompany focus-induced prominence that is marked by pitch accent in Germanic languages such as English, German, and Dutch (Cho, 2016; Cho & McQueen, 2005; de Jong, 1995; Fletcher, 2010; Mücke & Grice, 2014). Similarly, the strongly reduced form is in line with the prominence distribution that is reflected in the prosodic structure of a given utterance, in such a way that pre-focal units are reduced, presumably to enhance the contrast of the upcoming unit (e.g., Cho, et. al., 2013; de Jong 2004) in a perceptually relevant way (see Cangemi & Baumann, 2020, for related discussion). Thus, we hypothesized that such a segmental detail, be it also reflected in a suprasegmental (durational) feature, reflects the prominence-related prosodic structure, and would in turn indicate the information structure serving as a marker of the level of givenness (i.e., whether it is the mentioned or the other object), as was indicated by the use of a pitch accent in Experiment 1.

3.1. Method

3.1.1. Participants

Thirty-two participants who had not participated in Experiment 1 were recruited from the Prolific platform on the conditions that their first language must be German and that their age must be in the range between 18 and 40. The actual ages of participants ranged from 18 to 37 with a median of 27.5. Eight participants were female and 24 were male.

3.1.2. Materials and procedure

The procedure and visual materials were the same as in Experiment 1. Only the auditory stimuli were changed. The baseline stimuli from Experiment 1 (from the same speaker), which contained a moderately reduced form [habm], were used as the base for generating versions with full and strongly reduced forms of *habe*n using PSOLA, so that the pitch contour and the duration of the first overlapping [ha] part of *haben* were the same. The [ha] had a duration of about 90 ms, and the [m] of the reduced form [ham] was 85 ms long, whereas the [bən] part of the full form had a duration of 144 ms, as observed in the natural productions. Fig. 6 shows examples of these stimuli for the target object *Biene* (Engl., 'bee'). It is worth noting that these two forms of the auxiliary verb properly differed in their total duration because doing otherwise would introduce an unwanted asymmetry—i.e., keeping the total word duration the same between the full and reduced forms would make them sound as if they were produced at different speech rates. Because the total durational difference was due to a different number of segments ([habən] vs. [ham]), it could still be segmental in nature rather than suprasegmental. These forms were spliced into the *verum* and *contrast* conditions to replace the original form of [habm] in the baseline sentences. The splicing was done at positive-going zero crossings. Note that in the contrast condition (where the object names were contrastive), we did not remove the pitch accent on the object name because doing so would make the speaker sound unreliable in his prosody, possibly leading the listeners to disregard prosodic cues (Roettger & Franke, 2019). The splicing process did not result in any discernible discontinuities in the pitch contours of the spliced stimuli, as can be seen in Fig. 6.

3.2. Results

3.2.1. Reaction times

We carried out the same data preprocessing steps that we used in Experiment 1 (all scripts available at OSF). This time, two trials were rejected because the reaction time was much longer than what could be expected for a trial, indicating that there was a timing-related problem on those trials. Moreover, two participants' data were rejected because they did not use the mouse button to end trials properly for more than half

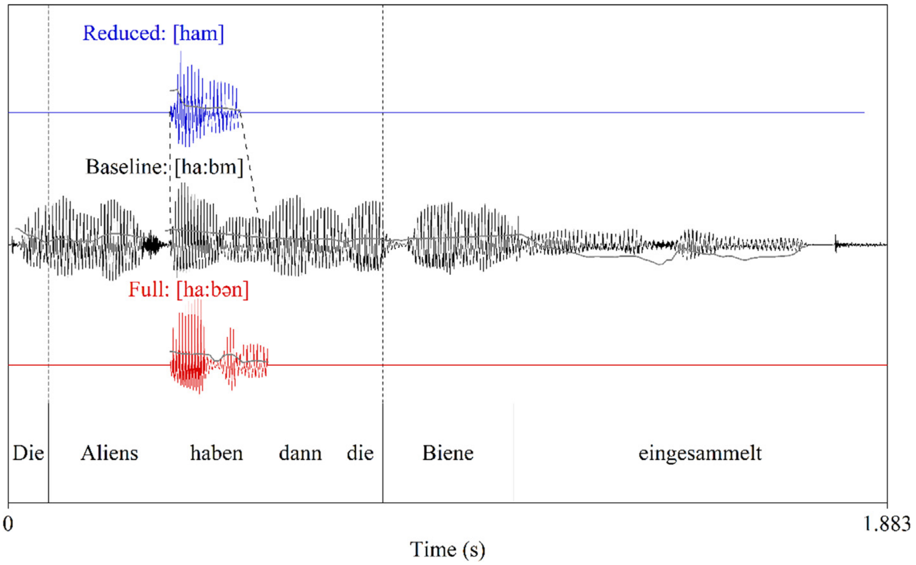


Fig. 6. Examples of stimuli as used in experiment 2. The top row shows the reduced version of the auxiliary, the middle row the baseline, typical version, and the bottom row the full version without any reduction. The dotted lines indicate how the reduced form was spliced into the sentence. Notice that the full form was of roughly similar duration as the default form.

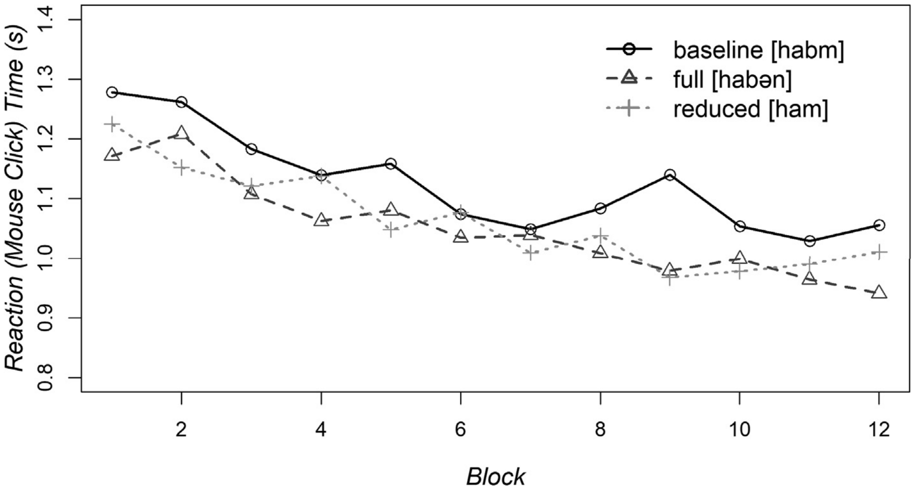


Fig. 7. Mean reaction (mouse click) times relative to the acoustic onset of the target word over the twelve blocks in Experiment 2 in which the auxiliary *haben* in the three conditions differed in terms of its segmental realization rather than its f0 realization—i.e., the baseline condition (with a moderately reduced form [habm]), the full condition (with no reduced form [habən]) matched to the *verum* focus condition in Experiment 1, and the strongly reduced condition ([ham]) matched to the *contrast* condition in Experiment 1.

the trials. In the remaining data set, 104 trials (=2.3%) were rejected for missing responses, and an additional 54 (=1.1%) were rejected because their reaction times had a normalized residual larger than absolute three in an intercept-only model that considered participant and item random effects. Fig. 7 shows the mean reaction (mouse click) time in the twelve blocks for the three conditions. The reaction time results indicate that participants clicked on the mouse button (to collect the object) earlier in the full and (strongly) reduced conditions than in the baseline (moderately reduced) condition, demonstrating an advantage for the conditions with the reduced and full forms compared with the baseline condition.

Table 3 shows the results of using a linear mixed-effects model with the maximally converging random effect structure on these data, with Block as the numeric predictor (range: 0–11) and two linearly independent contrasts for the three-level Condition predictor. As in Experiment 1, the first contrast compares the baseline condition (with the typical *habm* pronunciation, mapped onto −2/3) with the two conditions with a marked segmental detail (which we call Baseline vs. Information mapped onto 1/3). The second contrast compares the two potentially information-bearing conditions against each other (with the full condition mapped onto 0.5 and the reduced condition to −0.5). The maximal model that converged excluded correlations between random slopes and contained random slopes for both contrasts over participants but, over items, only the random slope for the contrast *Baseline vs. Informative* remained. Random slopes for Block and its interactions had to be removed.

Table 3
Results of the linear mixed-effects model for the reaction times of Experiment 2.

Predictor	B (SE)	t (df)	p
Intercept	1230 (57)	21.558 (33)	<0.001
Block	–22 (2)	−14.54 (4191)	<0.001
Condition			
Baseline vs. Information	−66 (20)	−3.268 (90)	0.002
whichInformation (Verum vs. Contrast)	6 (23)	0.246 (155)	0.806
Condition * Block			
Block:(Baseline vs. Information)	−1 (3)	−0.256 (4191)	0.798
Block:whichInformation	−3 (3)	−1.069 (4188)	0.285

Note: The syntax of the final model was: (1000*RT_targetOnset) ~ block*(baselineVsInformation + whichInformation) + (1 + baselineVsInformation + whichInformation||participant) + (1 + baselineVsInformation||item).

As summarized in Table 3, the results indicate an advantage for the two conditions with marked segmental detail compared to the baseline condition. In other words, participants reacted more quickly in both the full and reduced conditions compared to the baseline condition. However, in contrast to Experiment 1, these two conditions did not differ from each other. There was an effect of Block, indicating that reaction times sped up during the course of the experiment, but we found no Condition x Block interaction, indicating that the observed advantages of segmental information in the full and reduced conditions remained stable throughout the experiment.

3.2.2. Mouse-tracking data

First, we present the raw mouse positions over time in Fig. 8. The data were processed in the same way as in Experiment 1. All participants had a sufficient number (more than half) of trials with a mouse movement in the time window from −500 ms to around the onset of the target (set to zero). As in Fig. 4 for Experiment 1, Fig. 8a (the left panel) shows all the trials, and Fig. 8b (the right panel) shows the data after excluding 562 trials (11% of the data) for which the mouse was already in the target corridor from the beginning. The remaining trials showed a change in the mouse position moving toward the target corridor during the −500 to 2000 ms window, and those position changes were used to calculate decision times (from the time of the first mouse movement toward the target corridor to the time at which the mouse entered the corridor).

Fig. 9 shows how those decision times developed for the three conditions across the twelve blocks. The results of the statistical analysis are summarized in Table 4. The maximal model that converged excluded correlations between random slopes and contained random slopes for both contrasts over participants. Random slopes for Block and its interactions had to be removed. For the item random effect, only a random

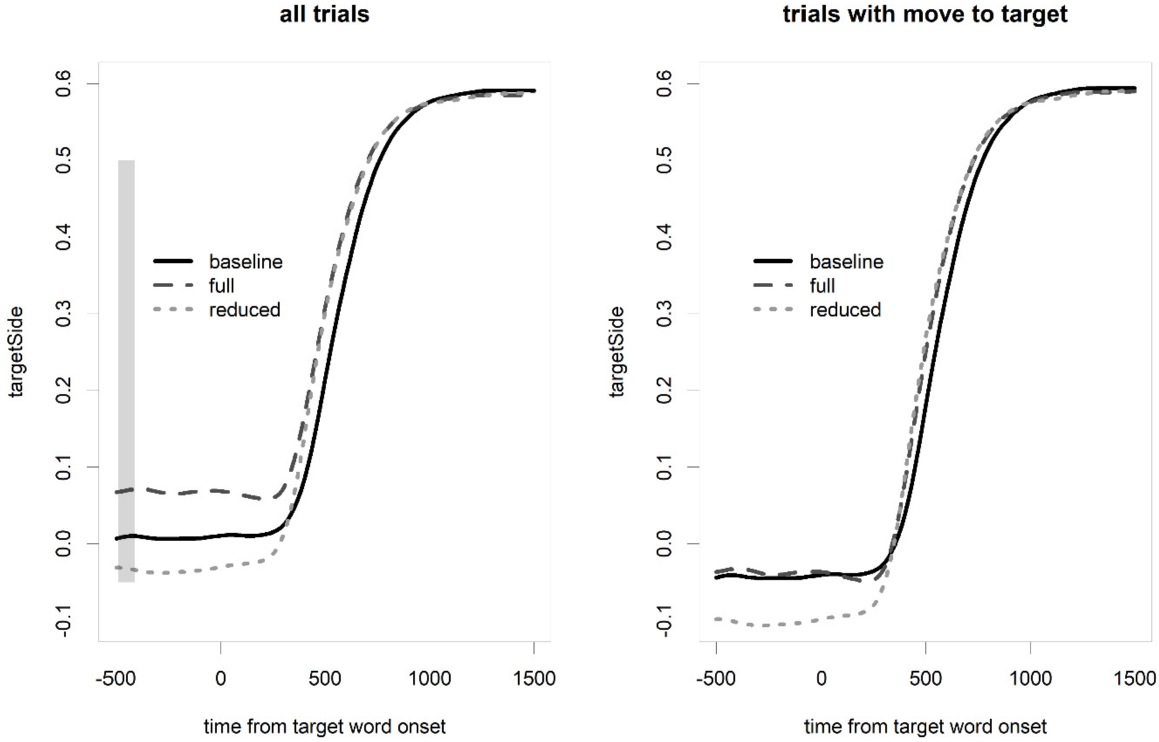


Fig. 8. Mouse positions in Experiment 2 in the time window from −500ams to 1500 ms relative to the target word onset. The left panel (a) shows all data and the right panel (b) only trials with a move towards the target after removing data (562 trials, 11%) in which the mouse position was always on the target side. The gray area on the left panel (a) indicates the range of endpoint of the auxiliary in the answer sentences relative to the onset of the target word.

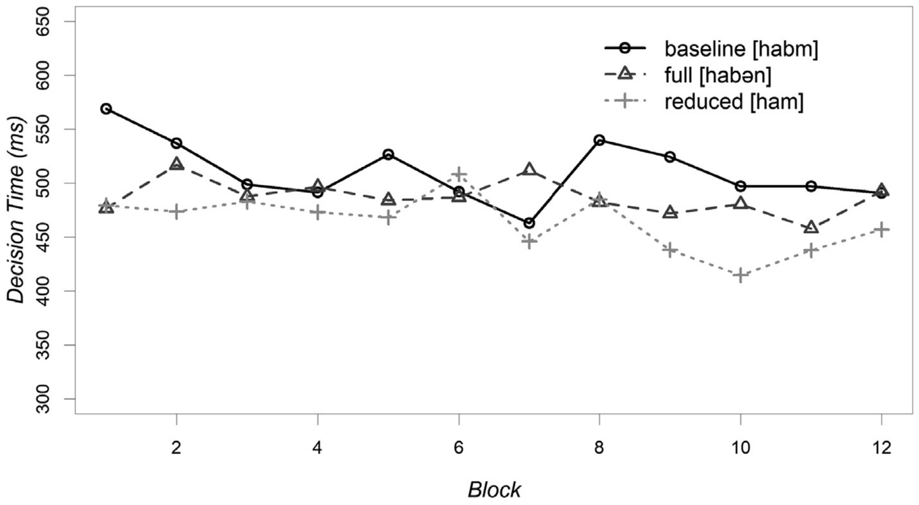


Fig. 9. Decision times in the mouse tracking data in Experiment 2. The decision times were calculated from the time at which the participant started moving the mouse cursor towards the target to the time at which the mouse entered the region (corridor) of the target.

intercept remained. As can be seen in Fig. 9, the results are basically similar to the results of the observed reaction times: the full and reduced conditions have an advantage over the baseline condition. We found no interaction between Condition and Block, as shown in Table 4, indicating that these effects were relatively stable throughout the course of the experiment (i.e., across blocks). It is worth noting, however, that the effects found in Experiment 2 appear to be much smaller than the ones found in Experiment 1, which can easily be seen by comparing Fig. 9 with Fig. 5, the figure showing the decision times in Experiment 1.

3.2.3. Discussion

The results show that the two conditions with potentially informative segmental detail have processing advantages over the baseline condition. That is, both reaction times (from the target onset to the mouse click to collect the object) and decision times (from the time of the mouse movement toward the target to the time of entering the target corridor) were faster for the target that had been mentioned in the full [habən] condition and for the target that was contrastive with the previously mentioned one in the reduced [ham] condition, relative to the baseline condition. The segmental hyperarticulation served

Table 4
Results of the linear mixed-effects model for the estimated decision times of Experiment 2.

	B (SE)	t (df)	p
Intercept	497 (32)	15.446 (36)	<0.001
Block	−3 (1)	−3.22 (4058)	0.001
Condition			
Baseline vs. Information	−44 (13)	−3.518 (237)	<0.001
whichInformation (Verum vs. Contrast)	10 (16)	0.593 (146)	0.554
Condition * Block			
Block:(Baseline vs. Information)	1 (2)	0.661 (4084)	0.509
Block:whichInformation	2 (2)	0.754 (4086)	0.451

Note: The syntax of the final model was: decision ~ block*(baselineVsInformative + whichInformative) + (1 + baselineVsInformative + whichInformative||participant) + (1 |item).

Table 5
The cross-experiment comparison of reaction times from Experiment 2 and 3. (Recall that the full form [habən] and the reduced [ham] conditions in Experiment 2 were paired with the carhon and the dog-bark conditions in Experiment 3, respectively.).

	B (SE)	t (df)	p
Intercept	1220 (44.5)	27.418 (62)	<0.001
Experiment	46 (82.3)	0.564 (21)	0.579
Block	−20 (2.2)	−8.927 (60)	<0.001
Condition			
Baseline vs. Information	−54 (14.5)	−3.761 (96)	<0.001
whichInformation (Verum vs. Contrast)	11 (16.3)	0.707 (266)	0.48
Experiment: Block	1 (4.1)	0.179 (92)	0.858
Exp * Condition			
Experiment: Baseline vs.Information	−29 (26.9)	−1.061 (379)	0.289
Experiment: whichInformation	−11 (32.1)	−0.347 (333)	0.728
Block * Condition			
Bock: Baseline vs. Information	−2 (1.8)	−0.829 (8014)	0.407
Block: whichInformation	−4 (2.1)	−1.915 (8013)	0.056
Exp * Block * Condition			
Experiment: Block: Baselinevs. Information	2 (3.7)	0.675 (8014)	0.5
Experiment: Block: whichInformation	1 (4.2)	0.285 (8016)	0.776

Note: the syntax of the final model was: (1000*RT_targetOnset) ~ expC*block * (baselineVsInformation + whichInformation) + (1 + expC + block + baselineVsInformation + whichInformation ||participant) + (1 + baselineVsInformation ||item).

as a cue to the information structure, just as the *verum* focus (which was expressed primarily through the suprasegmental feature of f0) did. It is also worth noting that the advantage was similar for the full and reduced forms compared to the baseline form, despite the reduced form being much shorter than the baseline form. This implies that although durational variation may arise with the segmental variation, it is not the main driving factor behind the results here.

Our results also suggest that segmental effects began early in the process, that is, from the decision time, just as we found for the effect of f0 in Experiment 1. (Recall that we employed two measures to test segmental effects—i.e., ‘decision time’ which was the time taken from the start of the mouse moving to the entrance into the target corridor *before* actually choosing the target object by activating the tractor beam, which temporally follows ‘reaction time’ which was taken to activate the tractor beam relative to the onset of the target *after* entering the target corridor.).

However, there are also clear differences from the effects observed in Experiment 1. First, the advantages are notably smaller, roughly half the size of those observed in Experiment 1. We interpret this as implying that suprasegmental features carry more robust cues than segmental details, as manipulated in Experiment 2, to the prominence-related prosodic structure that can be mapped onto the information structure. Second, the effect of both informative conditions (*verum* and *contrast*) is quite similar here, whereas in Experiment 1, the benefit of the *verum* condition was much larger than the benefit of the *contrast* condition. The benefit of the *verum* condition over the *contrast* condition in Experiment 1 might not be surprising because the difference in the pitch contours between the stimuli in the conditions of baseline and contrast stimuli in Roettger and Franke (2019) was small, with only a small rise on the initial subject noun to generate a “hat” pattern that is common in German sentences (Grice et al., 2017). In our case, it might be plausible that the pronounced segmental reduction of 'haben' in the contrast condition could potentially suggest an upcoming contrast more than the absence of an early f0 rise, which is usually characteristic of a “default” German sentence without a contrastive pitch accent.

Another difference between the two experiments was found in the time course of the advantage that the two informative conditions had over the baseline condition throughout the experiment. In Experiment 1, the benefits of the two experimental (*verum* and *contrast*) conditions increased as the experiment progressed (as evident in the significant Condition x Block interaction), which could partly be due to a task-specific learning effect. But in Experiment 2, the effects were relatively stable throughout the experiment (as evidenced by finding no Condition x Block interaction), with only a general speed-up of mouse clicks (reaction times) during the experiment (the Block effect).

However, despite the lack of a Condition x Block interaction during Experiment 2 (which implies no possible learning effect embedded in the observed results), we cannot rule out the possibility that some sufficient learning about the nature of the experiment had already taken place during the training block that left no further evidence during the main experiment. For instance, during the training block, participants might simply have learned the co-variation of the phonetic form and the eventual target (a full form → Go to the mentioned object; a reduced form → Go to the other (new) object). Note that such learning can occur without the participant being aware of it (Reed & Johnson, 1994). That possibility therefore undermines the validity of our interpretation that the observed effects reflect the listener's use of segmental detail as a cue to prosodic structure. We thus ran another experiment to better understand the nature of the results observed in Experiment 2.

4. Experiment 3

In Experiment 3, we tested whether the effects observed in Experiment 2 indeed reflected the listener's use of segmental detail to infer the information structure of the dialogue by computing the prosodic structure of the utterances or reflected a simple learning process during the task. In this experiment, participants had the same chance to learn the co-variation between the auditory stimulus and the target, but this time without a relation between the segmental detail and prosodic structure. That is, we used the same material as in Experiment 2, but we masked the phonetic form of the auxiliary in a consistent way in the two potentially informative conditions (reduced and full) but left it unmasked in the baseline condition. The full form was masked by a brief sound of a car horn, and the reduced form was masked by a brief sound of a dog bark. Therefore, the same possibility for learning existed as in Experiment 2 (a car horn → Go to the mentioned object, a dog bark → Go to the other (new) object), but the naturally driven relationship between this cue and the target (as in the relation between the full form and the prosodic weight) was removed. If the results of Experiment 2 were due to participants learning the existing co-variation between the segmental detail and the prosodic weight (prominence), the participants should also learn the co-variation between the sound type (a car horn versus a dog bark) and the referent type (previously mentioned or not).

4.1. Method

4.1.1. Participants

Thirty-three participants who had not participated in either of the two earlier experiments were recruited from the Prolific platform on the conditions that their first language must be German and that their age must be in the range between 18 and 40. The actual ages of the participants ranged from 18 to 40 with a median of 26. Six participants were female and 27 were male.

4.1.2. Materials and procedure

The materials were the same as in Experiment 2, with the additional masking of the auxiliary verb. To that end, we used a brief car-horn sound and a brief dog-bark sound found in publicly available sound databases. They were cut to the length—cutting done at positive zero crossings—of the respective durations of the auxiliary verb forms (full form [habən] and reduced form [ham]). The amplitude was adjusted so that the signal-to-noise ratio was -12db (i.e., the masking sound was about 4 times louder than the speech sound), to ensure a proper masking of the phonetic form of the auxiliary *haben*.

4.2. Results

4.2.1. Reaction times

After preprocessing the data in the same way as in Experiment 2 (all scripts are available at OSF), 17 trials (0.39% of the data) were rejected because the reaction was longer than the trial should have been. Moreover, two participants' data were rejected because they did not use the mouse button to end trials properly for more than half the trials. In the remaining data set, 191 trials (=4.6%) were rejected for missing responses, and an additional 42 (=0.95%) were rejected because their reaction times had a normalized residual larger than absolute three in an intercept-only model that considered participant and item random effects.

Fig. 10 shows the mean reaction times in the twelve blocks for the three conditions. We see a general decrease in the reaction times and an advantage for the two masking conditions over the baseline conditions. This is similar in Experiment 2. Since the main question is whether the results differ between the experiments, we immediately make a cross-experiment comparison (rather than falsely assuming that a significant effect in one experiment and the absence thereof in another constitutes a difference between the two experiments). The predictors were coded along the lines of the earlier experiments. Block was used as the numeric predictor (range: 0–11) and two contrast-coded predictors for the Condition variable, using the analogue coding as in Experiment 2. The first contrast compares the two conditions with masking (Exp 3) or marked phonetic form (Exp 2) of the auxiliary (mapped onto 1/3) with the baseline condition (mapped onto −2/3), and the second contrast compares the two potentially informative conditions with masking/marked phonetic forms against each other (with the full condition mapped, masked with a car horn in Exp. 3, onto 0.5 and the reduced condition, masked with a dog bark in Exp. 3, mapped onto −0.5). Experiment was contrast coded with Experiment 3 mapped onto −0.5 and Experiment 2 onto 0.5, reflecting the fact that Experiment 2 is potentially more informative.

The maximal converging model had random slopes of Experiment, Block, and both condition contrasts over participants but only a random slope of the baseline vs. informative conditions contrast over items. The results for reaction times (Table 5) show only an overall advantage for the two experimental conditions over the baseline condition. The interaction of those effects with Experiment and the three-way interaction with Experiment and Block failed to reach significance. Thus, as far as the results on reaction times are concerned, the two experiments appear to have no meaningful differences.

4.2.2. Mouse-tracking data

The data were processed in the same way as in Experiments 1 and 2. Fig. 11 shows the raw mouse positions over

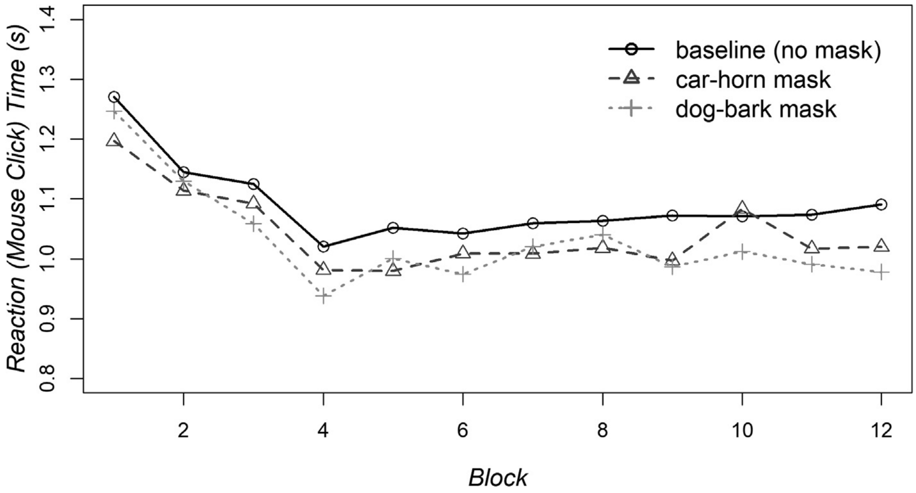


Fig. 10. Mean reaction (mouse click) times relative to the acoustic onset of the target word over the twelve blocks in Experiment 3 in which [habən] (the full form in Experiment 2) was masked by a car-horn sound and [ham] (the reduced form in Experiment 2) by a dog-bark sound.

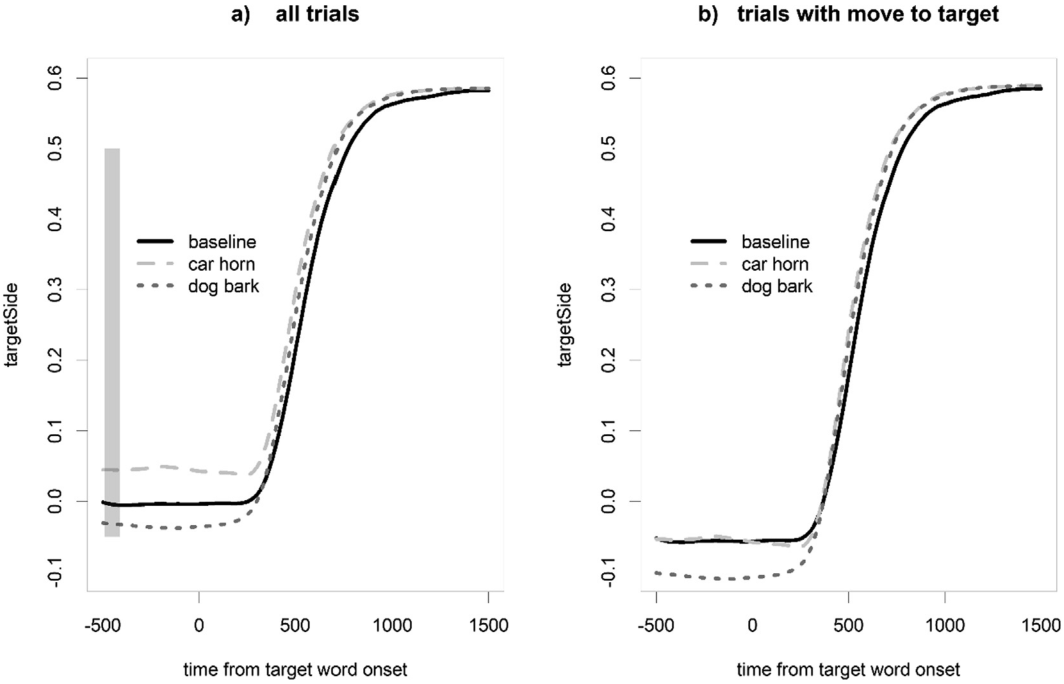


Fig. 11. Mouse positions in Experiment 3 in the time window from −500ams to 1500 ms relative to the target word onset. [habən] (the full form in Experiment 2) was masked by a car-horn sound and [ham] (the reduced form in Experiment 2) by a dog-bark sound. The left panel (a) shows all data and the right panel (b) only trials with a move towards the target after removing data (523 trials, 11.9%) in which the mouse position was always on the target side. The gray area on the left panel (a) indicates the range of endpoint of the auxiliary in the answer sentences relative to the onset of the target word.

time for all the trials (a) and after removing data for 523 trials (11.9%) in which the mouse position was always on the target side (b). All participants had a sufficient number of trials useful for calculating decision times (mouse movements in more than half the trials during the window from −500 ms and 2000 ms with the target onset set to zero). As can be inferred from Fig. 11b, the masked conditions, which contain potential cues (a car-horn sound and a dog-bark sound) to the target (if learned during the experiment), had only a subtle advantage over the baseline condition.

Fig. 12 shows how decision times evolved for the three conditions across the twelve blocks, showing an advantage for the two masking conditions that arose over the course of the experiment. Note that this pattern differs from what was observed in Experiment 2, where the advantage of the two experimental conditions over the baseline condition remained relatively stable throughout the experiment. We investigated whether this difference is genuine rather than merely a result of a learning effect that might have arisen during the experiment. The predictors were coded in the same way as in the reaction time analysis above. The maximally converging model for decision times across experiments included random slopes for Block and both condition contrasts over participants, with only a random intercept for items. The results revealed significant differences between the two experiments. As summarized in Table 6, there was a significant difference between

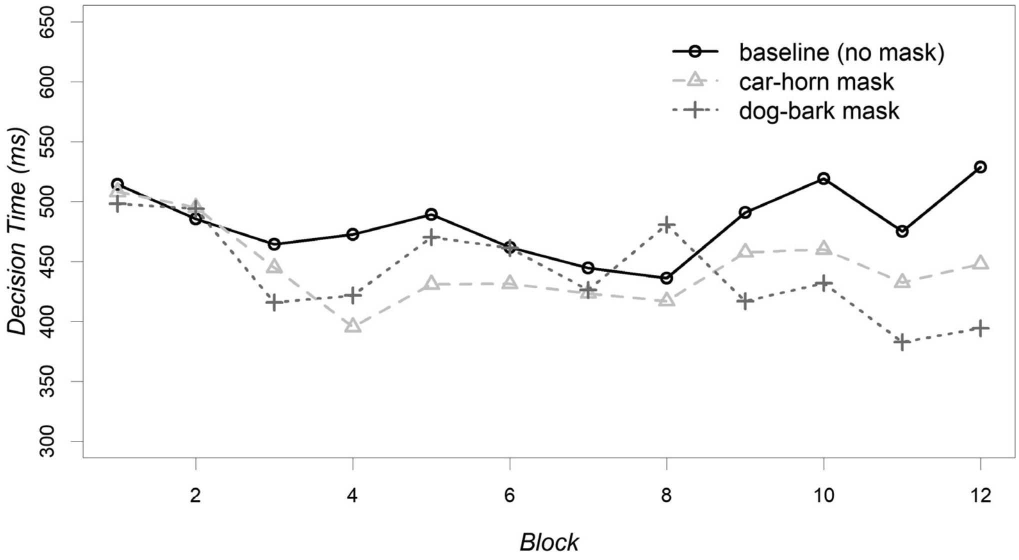


Fig. 12. Decision times in the mouse tracking data in Experiment 3. [habən] (the full form in Experiment 2) was masked by a car-horn sound and [ham] (the reduced form in Experiment 2) by a dog-bark sound. The decision times were calculated from the time at which the participant started moving the mouse cursor towards the target to the time at which the mouse entered the region (corridor) of the target.

Table 6
The cross-experiment comparison of decision times from Experiment 2 and 3. (Recall that the full form [habən] and the reduced [ham] conditions in Experiment 2 were paired with the car-hon and the dog-bark conditions in Experiment 3, respectively.).

	B (SE)	t (df)	p
Intercept	506 (25)	20.233 (71)	<0.001
Experiment	13 (27.1)	0.474 (768)	0.636
Block	−3 (1.4)	−2.18 (65)	0.033
Condition			
Baseline vs. Information	–32 (10.4)	−3.119 (233)	0.002
whichInformation (Verum vs. Contrast)	15 (12)	1.224 (315)	0.222
Experiment: block	−1 (2.7)	−0.556 (93)	0.579
Exp * Condition			
Experiment: Baseline vs. Information	−26 (20.5)	−1.259 (282)	0.209
Experiment: whichInformation	−13 (23.8)	−0.552 (372)	0.581
Block * Condition			
Bock: Baseline vs. Information	−2 (1.4)	−1.21 (7791)	0.226
Block: whichInformation	−1 (1.6)	−0.746 (7792)	0.456
Experiment * Block * Condition			
Experiment: block: Baseline vs. Information	6 (2.8)	2.024 (7796)	0.043
Experiment: Block: whichInformation	6 (3.3)	1.807 (7795)	0.071

Note: the syntax of the final model was: decision ~ expC*block*(baselineVsInformation + whichInformation) + (1 + block + baselineVsInformation + whichInformation ||participant) + (1|item).

the two experimental conditions and the baseline condition, but crucially, there was a significant three-way interaction of Condition with Experiment by Block for the contrast comparing the potentially informative conditions with the baseline condition. To investigate the source of this interaction, we used the function *emtrends* from the package *emmeans* (Lenth et al., 2023) to calculate the slope of block (i.e., the extent to which participants get faster over the course of the experiment) for the baseline and experimental conditions in Experiment 2 and Experiment 3. The function also supplies confidence intervals for these slopes. Fig. 13 shows the results of this procedure and elucidates the source of the three-way interaction between Block, Condition, and Experiment. In Experiment 2, there is a general speed-up of responses over the course of the experiment. This is not unusual to observe. Participants may simply get more comfortable with the procedure and tend to get faster over the course of an experiment. However, in Experiment 3, the confidence intervals for the baseline and masking conditions exclude the other mean, indicating that learning in the masking conditions is stronger than in the baseline condition. That is, the results indicate that participants learn over the course of the experiment that the masking sounds reliably indicate the target object.

These results suggest that learning could be a potential confounding factor in studies involving the use of prosody for information structure. Participants might learn covariances within the experimental context and then apply these learnings to enhance their performance in the given task. This implies that a study which finds that a certain prosodic feature is used to predict the progression of a sentence requires further validation. This is necessary to eliminate the possibility that any observed effect may be attributed to the prosodic feature's predictive role within that specific study. In other words, it is crucial to assess whether participants are merely responding to the contingencies within the experiment, as opposed to applying knowledge acquired outside the laboratory setting. However,

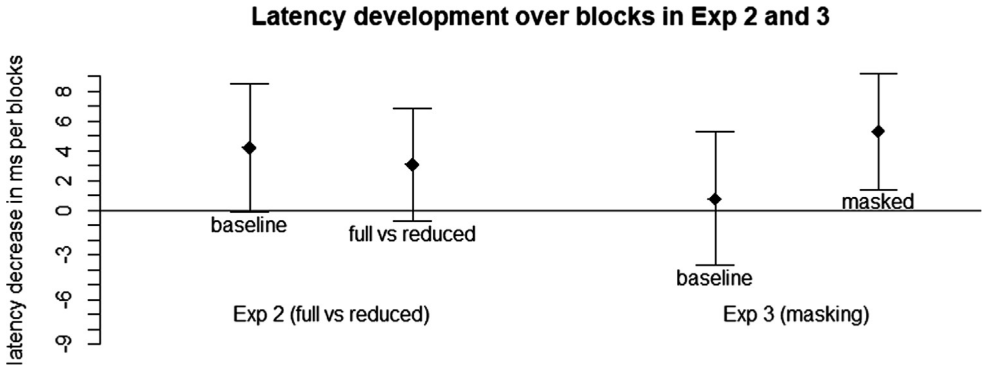


Fig. 13. Learning effects in Experiments 2 and 3, for both (potentially) informative and baseline conditions. Note that as latency decreases over blocks, higher y-axis values indicate greater latency decrease and, therefore, more learning.

the three-way interaction between Experiment, Condition, and Block found for the decision times above shows that this is an unlikely explanation for the current results, indicating that segmental effects observed in Experiment 2 cannot be simply due to a learning effect developed during the experiment. The results of Experiment 2 therefore likely reflect the use of segmental information as a proxy for prosodic information related to information structure.

5. General discussion

In this study, we have investigated whether a segmental detail can be used as a cue to the prosodic structure, which interacts with other linguistic structures such as syntax and information structure (e.g., Beckman, 1996; Elfner, 2018; Röhr et al., 2022; Shattuck-Hufnagel & Turk, 1996; Zahner-Ritter et al., 2022). To test this, we focused on the relationship between prosody and information structure, which, in German, can lead to prosodic prominence for the auxiliary verb (*haben*, ‘to have’) in an utterance that confirms the answer to a yes/no question (a case of the so-called *verum* focus, see Turco et al., 2014). An earlier study by Roettger & Franke (2019) had indeed shown that listeners could use pitch (suprasegmental) cues related to the *verum* focus to infer the information structure of utterances exchanged among interlocutors, especially to infer whether an upcoming referent in the answer to a question would be given or new. In this study, we explored the listener’s use of a segmental detail in computing the prosodic structure of a given utterance.

Testing segmental effects is particularly important because segmental detail is generally considered to provide cues about the ‘what’ component of a linguistic message (e.g., segmental and lexical information in a non-tonal language), whereas the ‘how’ component is manifested primarily in the domain of suprasegmental features, often in conjunction with the speaker’s production of a particular prosodic structure in a given utterance. However, given that segmental details are also systematically modulated by the prosodic structure of a given utterance (Cho et al., 2017; Jang et al., 2018, 2023; Keating et al., 2003; Mücke & Grice, 2014; Pouplier, 2022), they should also contribute to constructing the how component in some way. In this context, we do not rule out the possibility that segmental details associated with the reduced versus full forms of any lexical item (the auxiliary verb 'haben' in this case) may also involve some durational modification. Therefore, the listener's interpretation of the segmental details might be influenced at least in part by the durational difference. What is important here is that segmentally different forms, whether reflected in duration or not, is utilized to reference prosodic structure at the utterance level. We therefore hypothesized that segmental details would also be exploited by the listener in computing the prosodic structure, and thus contribute to listener understanding of the information structure of a given utterance, even in the absence of suprasegmental cues. The segmental detail tested in this study compared the coarticulated segmental reduction and hyperarticulation of the auxiliary verb *haben*, varying from [ham] (reduced) to [habm] (moderately reduced) to [habən] (full/hyperarticulated). The results we obtained from a series of experiments (Experiments 1–3) indeed lend support to our hypothesis—i.e., with the segmental detail alone, listeners were able to predict an upcoming referent according to the information structure, indicating that they used the segmental detail in computing a prosodic structure, similar to the way that they use suprasegmental (pitch) cues, though the magnitude of the effects differed. In the remainder of this section, we recapitulate specific findings and offer some theoretical implications for the convoluted interplay of phonetics, prosodic structure, and other higher-order linguistic structures.

Experiment 1 demonstrated that a remotely accessed online mouse-tracking paradigm can yield similar effects to an earlier lab-based version (Roettger & Franke, 2019). The results indicated that listeners used prosodic information to infer information structure. In Experiment 2, listeners exploited prosodically conditioned segmental details of the auxiliary verb 'haben,' even in the absence of f0 information, similar to the way the pitch accent was used. However, based on Experiment 2, we could not rule out that these effects were due to learning during the experiment (i.e., hearing a full form led to selecting the mentioned object, while hearing a reduced form led to selecting the other object).

Experiment 3 aimed to evaluate this possibility by systematically masking the auxiliary verb using different noises. One type of masking noise (a car horn) was always paired with the full form, while another type (a dog bark) was always paired with the reduced form. This allowed for the same kind of learning as in Experiment 2. However, listeners could not exploit this co-varying relationship between masking noises and target objects at the beginning of the experiment, and the predictive effect only developed over time. Interestingly, the effect eventually became stronger than the effect of segmental detail observed in Experiment 2, indicating robust learning during the course of the experiment. In contrast, the effect of segmental detail observed in Experiment 2 was significant from the very beginning and remained stable throughout the experiment. These differences, when considered in a cross-experiment comparison, suggest that the results of Experiment 2 cannot be attributed to mere learning.

This finding supports our hypothesis that the prominence-related prosodic weight the speaker put on the auxiliary verb was phonetically encoded on the segmental phonetic detail with the degree of segmental reduction due to coarticulation and segmental hyperarticulation, which is in turn available to the listener to compute the prominence structure, the source of prosodic weight imposed on the auxiliary verb.

It is also worth noting that in Experiment 2, the segmental effects were more stable (with no obvious Condition by Block interaction) not only relative to the effects observed in Experiment 3 (with non-linguistic masking sounds) but also to those in Experiment 1 (with pitch accents). In the latter two cases, we found Condition and Block interactions, with some evidence that the effects of Condition grew larger as the experiment progressed through its 12 blocks. Those effects indicate that as the experiment continued, participants performed better, presumably by learning (becoming aware of) the co-variation between the stimulus and the target noun—i.e., the linguistically licensed relationship between pitch accents and potential targets (Experiment 1) or the unnatural (ecologically invalid) relationship between non-linguistic noises and potential targets (Experiment 3). Note again that Experiments 1 and 3 still differed from each other in that the effect was evident from the very beginning for the linguistically relevant relationship, but not for the unnatural relationship. This also means that participants indeed made use of the prosodic features (pitch accent) in computing prosodic structure, independent of learning, although a learning effect could facilitate the task during the course of the experiment. It is reasonable to question why no such learning effect was evident for the segmental detail in Experiment 2. That is not an easy question to answer, but what appears to be the case is that learning about segmental phonetic variants of the auxiliary *haben* and their relations to targets is more cumbersome than learning about pitch accent and non-linguistic noises. Of course, this possibility leads to another question, which is why listeners are not good at noticing segmental phonetic variation (cf. Ernestus, 2013), despite the ubiquitous use of such forms and their apparent ability to consciously perceive speech as a stream of segments (which might be reinforced by reading experience, Morais et al., 1979). We suggest that the lack of a learning effect might have to do with the differential auditory-perceptual saliency associated with the stimuli. Both the pitch accent with a substantial f0 rise used in Experiment 1 and the masking noise with an outstanding increase in intensity used in Experiment 3 are likely to have more auditory-perceptual impacts than the segmental detail. It is therefore reasonable to assume that although listeners can use explicit strategies with auditory-perceptual pitch accents and different masking noises, learning about segmental phonetic variation might rely on only implicit learning processes, which are arguably slower than explicit strategies (cf., Cadierno et al., 2020). During the short time of Experiment 2, such learning with segmental phonetic variation might not have taken place. (As an anonymous reviewer pointed out, the observed lack of a learning effect may also be attributed to the possibility that reduction processes are harder to learn in this particular context, especially when the full form [habən] is an extremely low-frequency form.).

Finally, it was noticeable that the magnitude of the effect of the segmental detail in Experiment 2 was smaller than that of the pitch accent information signaled by the suprasegmental feature of f0 observed in Experiment 1. As we discussed above, it is conceivable that the difference is phonetically grounded, at least in part stemming from differential auditory-perceptual impacts carried by conspicuous pitch accents versus segmental phonetic variation in the absence of a pitch accent. Seen from a different angle, the difference can be taken as *phonologically* grounded as well, conditioned by the language's grammar of intonational phonology. The theoretical premise of intonational phonology, which is generally rooted in the Autosegmental Metrical (AM) theory (Beckman & Pierrehumbert, 1986; Grice et al., 2005; Ladd, 2008), is that a prosodic structure, especially the tune of a given utterance, is constructed by bridging gradient physical f0 events and systematically organizing patterns by mapping the phonological representations of assigned underlying tones to continuous f0 events. In this theoretical framework, the *verum* focus-related pitch accent (in Experiment 1) can be seen as being realized by a phonologically-defined tonal target (e.g., L*+H) to the test word *haben*, governed by the intonational grammar of the language. Thus, in the absence of the primary phonological (f0) features of prominence, the phonological representations might not be easily retrieved based on the (secondary) segmental detail alone, independent of the assumed differential auditory-perceptual impacts carried by an available cue. This might not differ from the case with segmental categories, which are often associated with one major acoustic cue, but they can also be influenced by secondary cues, especially if the primary cue is not entirely clear (Repp, 1983).

Still, it is important to reiterate that, based on the available segmental detail alone, listeners can infer the information structure of utterances by referencing prosodic structure. However, it is possible to conceive of an alternative account in which speakers select different versions of 'haben' in production to signal information structure directly (independent of prosody), and listeners might then learn to associate the different segmental realizations of 'haben' with different information structures, with no reference to prosodic structure. We do acknowledge that such effects, independent of prosody, have been observed with regard to the impact of predictability and repetition on word durationt (Baker & Bradlow, 2009). For the case of phrase-level prominence with a pitch accent, however, there is ample evidence that segmental hyperarticulation versus reduction, which is also reflected in duration, comes about as a direct consequence of the presence or absence of a pitch accent. Given this relationship between segmental realization and prominence (with pitch accent associated with verum focus in this case), we suggest that assuming another, parallel mechanism is not particularly parsimonious. However, what remains an open question is whether the different forms of reductions of 'haben' are a consequence of shortening caused by a prosodic-structurally driven phonetic reduction process or whether different forms of *haben* are represented lexically and there is a prosodic-structurally conditioned choice of word forms. (Note that the reduction of /bən/ to [bm] is a common process in German, which applies to other words as well.).

If we assume therefore that the change in form of *haben* is prosodically conditioned, this has some theoretical implications for speech perception, especially with regard to the so-called *prosodic analysis* model, which refers not to a particular theory, but to a general view adopted by some researchers (e.g., Salverda et al., 2003; Cho, et al., 2007; Mitterer et al., 2019; McQueen & Dilley, 2020; Steffman, 2021; Steffman et al., 2022). The crux of this view is that prosodic information that signals the prosodic structure of a given utterance (for the how component of a linguistic message) is processed by a 'Prosodic Analyzer' (cf., Cho et al., 2007) in parallel with a segmental analysis (for the what component), so that computing an alignment of segmental information and prosodic structure modulates lexical access. But our results, together with evidence coming from speech production studies (e.g., Keating et al., 2003; Mücke & Grice, 2014; Cho et al., 2017, Jang et al., 2018, 2023), suggest that prosodic analysis is convoluted with segmental analysis in some principled ways, indicating no strong division between prosodic and segmental analyses, as the prosodic analysis model seems to assume. Further studies to integrate segmental details in computing a prosodic structure are certainly needed to refine the theoretical approaches that adopt the view of the prosodic analysis model. One possible avenue to embark on such a study could pertain to the time course for processing segmental versus prosodic information, which can be tested using a visual world paradigm of eye-tracking.[6] The prosodic analysis model generally predicts that prosodic information is processed relatively late in a post-lexical stage (Kim et al., 2018; McQueen & Dilley, 2020; Mitterer et al., 2019), whereas segmental information might be processed almost immediately (Mitterer & Reinisch, 2013; Reinisch & Sjerps, 2013; Toscano & McMurray, 2015). It will thus be particularly interesting to test whether a segmental detail that carries information for both the what and how components of the linguistic message is processed at an early stage and later integrated with prosodic information to compute the prosodic structure, or whether the processing of the segmental detail related to the how component is delayed to a later stage and used along with prosodic information to compute the prosodic structure.

6. Conclusion

In this study, we used a remotely accessed online version of a mouse-tracking experimental paradigm and collected interpretable data about the effects of both suprasegmental and segmental details in speech perception. The results of three experiments suggest that listeners use not only the primary phonological features of pitch accent (reflected in f0 realization), but also segmental detail in inferring the information structure of exchanged utterances, reflecting the interplay between prosodic and information structures. It is evident that listeners could predict whether an upcoming referent noun would be the one given in the preceding context or new based on the phonetic form of the auxiliary verb *haben* that preceded the target noun. In interpreting the results, we have explained that our findings must be understood in relation to the auditory-perceptual phonetic impacts of the stimuli (pitch accent–driven versus segmental detail) and the phonological impacts of the primary tonal features assigned by the grammar of intonational phonology, as well as a possible learning effect that this kind of experiment may not be entirely free from. The emerging evidence implies that the segmental detail should be integrated with the suprasegmental cues in computing the prosodic structure of a given utterance. Such integration may come later over the course of speech processing when considering its nature of post-lexical processing, but earlier when considering the fact that segmental details are generally processed early on. It remains to be seen how the current findings can be incorporated into theories of speech comprehension that attempt to take into account the convoluted relationship between segmental and suprasegmental features or between the what and how components of the linguistic message.

CRediT authorship contribution statement

Holger Mitterer: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Sahyang Kim:** Writing – review & editing, Validation, Methodology, Funding acquisition, Data curation, Conceptualization. **Taehong Cho:** Writing – review & editing, Writing – original draft, Validation, Methodology, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank all the online participants. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884).

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439. https://doi.org/10.1006/jmla.1997.2558.

Baker, R. E., & Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech, 52*(4), 391–413. https://doi.org/10.1177/0023830909336575.

Beckman, M. E. (1996). The parsing of prosody. *Language and Cognitive Processes, 11* (1–2), 17–68. Scopus.

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

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*, 341–345.

Cadierno, T., Hansen, M., Lauridsen, J., Eskildsen, S., Fenyvesi, K., Hannibal Jensen, S., & Wieschen, M. (2020). Does younger mean better? Age of onset, learning rate and shortterm L2 proficiency in young Danish learners of English. *Vigo International Journal of Applied Linguistics, 57–86*. https://doi.org/10.35869/vial.v0i17.1465.

[6] We tried to run an online eye-tracking version of Experiment 1, but we were disappointed with the data quality, with the maximal fixation proportions on the targets not going much above 50% (compared with 90% in lab-based tasks).

Cangemi, F., & Baumann, S. (2020). Integrating phonetics and phonology in the study of linguistic prominence. *Journal of Phonetics, 81*, 100993. https://doi.org/10.1016/j.wocn.2020.100993.

Cho, T. (2016). Prosodic boundary strengthening in the phonetics-prosody interface. *Language and Linguistics Compass, 10*(3), 120–141. https://doi.org/10.1111/lnc3.12178. Scopus.

Cho, T. (2022). Linguistic functions of prosody and its phonetic encoding with special reference to Korean. In K. Horie, Y. Akita, D. Y. Kubota, & A. Utsugi (Eds.), *Japanese/Korean linguistics 29* (pp. 1–24). CSLI Publications.

Cho, T., McQueen, J., & Cox, E. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics, 35*(2), 210–243.

Cho, T., Kim, J., & Kim, S. (2013). Preboundary lengthening and preaccentual shortening across syllables in a trisyllabic word in English. *Journal of the Acoustical Society of America, 133*(5), EL384–EL390.

Cho, T., Kim, D., & Kim, S. (2017). Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *Journal of Phonetics, 64*, 71–89. https://doi.org/10.1016/j.wocn.2016.12.003.

Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics, 33*(2), 121–157. https://doi.org/10.1016/j.wocn.2005.01.001.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language, 47*(2), 292–314. https://doi.org/10.1016/S0749-596X(02)00001-3.

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America, 97*(1), 491–504. https://doi.org/10.1121/1.412275.

de Jong, K. (2004). Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *Journal of Phonetics, 32*, 492–516.

Elfner, E. (2018). The syntax-prosody interface: Current theoretical approaches and outstanding questions. *Linguistics Vanguard, 4*(1). https://doi.org/10.1515/lingvan-2016-0081.

Ernestus, M. (2013). Halve woorden [Inaugural lecture]. *Radboud University.* https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_1703313_6.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. The handbook of phonetic sciences. *Blackwell Handbooks in Linguistics*, 523–602. Scopus.

Garallek, M. (2013). *Production and perception of glottal stops [dissertation]*. UCLA.

Grice, M., Baumann, S., & Benzmüller, R. (2005). German Intonation in Autosegmental-Metrical Phonology. In S.-.-A. Jun (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 55–83). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199249633.003.0003.

Grice, M., Ritter, S., Niemann, H., & Roettger, T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics, 64*, 90–107. https://doi.org/10.1016/j.wocn.2017.03.003.

Hartshorne, J. K., de Leeuw, J. R., Goodman, N. D., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods, 51*(4), 1782–1803. https://doi.org/10.3758/s13428-018-1155-z.

IPDS (Director). (1994). *The Kiel Corpus of Spontaneous Speech*. Universität Kiel.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language, 58*(2), 541–573. https://doi.org/10.1016/j.jml.2007.06.013.

Jang, J., Kim, S., & Cho, T. (2018). Focus and boundary effects on coarticulatory vowel nasalization in Korean with implications for cross-linguistic similarities and differences. *The Journal of the Acoustical Society of America, 144*(1), EL33–EL39. https://doi.org/10.1121/1.5044641.

Jang, J., Kim, S., & Cho, T. (2023). Prosodic structural effects on non-contrastive coarticulatory vowel nasalization in L2 English by Korean learners. *Language and Speech*, 00238309221108657. https://doi.org/10.1177/00238309221108657.

Keating, P. A. (2006). Phonetic encoding of prosodic structure. In *Speech production: Models, phonetic processes and techniques* (pp. 167–186). Psychology Press. https://doi.org/10.1016/B0-08-044854-2/00030-4.

Keating, P. A., Cho, T., Cecile, F., & Hsu, C. (2003). Domain-initial strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.). *Phonetic Interpretation: Papers in Laboratory Phonology* (Vol. 6, pp. 145–163). Cambridge University Press.

Kehrein, W., & Golston, C. (2004). A prosodic theory of laryngeal contrasts. *Phonology, 21*(3), 325–357. https://doi.org/10.1017/S0952675704000302.

Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America, 134*(1), EL19–EL25.

Kim, S., Mitterer, H., & Cho, T. (2018). A time course of prosodic modulation in phonological inferencing: The case of Korean post-obstruent tensing. *PLOS ONE, 13*(8), e0202912.

Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language, 40*(2), 153–194. https://doi.org/10.1006/jmla.1998.2620.

Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication, 34*(4), 391–405. https://doi.org/10.1016/S0167-6393(00)00058-3.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmerTest'. *R Package Version, 2*.

Ladd, D. R. (2008). *Intonational Phonology* (2nd ed.). Cambridge University Press. Doi: 10.1017/CBO9780511808814.

Lehiste, I. (1970). *Suprasegmentals.* MIT Press.

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (1.8.9) [Computer software]. https://cran.r-project.org/web/packages/emmeans/index.html.

Li, H., Kim, S., & Cho, T. (2020). Prosodic structurally conditioned variation of coarticulatory vowel nasalization in Mandarin Chinese: Its language specificity and cross-linguistic generalizability. *The Journal of the Acoustical Society of America, 148*(3). https://doi.org/10.1121/10.0001743.

Lohnstein, H. (2016). Verum Focus. In C. Féry & S. Ishihara (Eds.), *The Oxford Handbook of Information Structure* (pp. 290–313). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642670.013.33.

McQueen, J. M., & Dilley, L. C. (2020). Prosody and spoken-word recognition. *The Oxford Handbook of Language Prosody*, 509–521.

Mitterer, H., Cho, T., & Kim, S. (2016). How does prosody influence speech categorization? *Journal of Phonetics, 54*, 68–79. https://doi.org/10.1016/j.wocn.2015.09.002.

Mitterer, H., Kim, S., & Cho, T. (2019). The glottal stop between segmental and suprasegmental processing: The case of Maltese. *Journal of Memory and Language, 108*, 104034. https://doi.org/10.1016/j.jml.2019.104034.

Mitterer, H., Kim, S., & Cho, T. (2021a). Glottal stops do not constrain lexical access as do oral stops. *PLOS ONE, 16*(11), e0259573.

Mitterer, H., Kim, S., & Cho, T. (2021b). The role of segmental information in syntactic processing through the syntax-prosody interface. *Language and Speech, 64*(4), 962–979. https://doi.org/10.1177/0023830920974401.

Mitterer, H., & Reinisch, E. (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. *Journal of Memory and Language, 69*(4), 527–545. https://doi.org/10.1016/j.jml.2013.07.002.

Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition, 7*, 323–331.

Mücke, D., & Grice, M. (2014). The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation? *Journal of Phonetics, 44*, 47–61. https://doi.org/10.1016/j.wocn.2014.02.003.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y.

Pouplier, M. (2022). Advancements of phonetics in the 21st century: a critical appraisal of time and space in articulatory phonology. *Journal of Phonetics, 95*.

R Development Core Team. (2022). *R: A language and environment for statistical computing* (4.2.1) [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org.

Redi, L., & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics, 29*(4), 407–429. https://doi.org/10.1006/jpho.2001.0145.

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(3), 585–594. https://doi.org/10.1037/0278-7393.20.3.585.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics, 41*(2), 101–116.

Repp, B. H. (1983). Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. *Speech Communication, 2*(4), 341–361. https://doi.org/10.1016/0167-6393(83)90050-X.

Roettger, T. B., & Franke, M. (2019). Evidential Strength of Intonational Cues and Rational Adaptation to (Un-)Reliable Intonation. *Cognitive Science, 43*(7), e12745.

Roettger, T. B., & Franke, M. (2022, February 17). *Dynamic speech adaptation to unreliable cues during intonational processing*. OSF. https://osf.io/xbh5m/.

Röhr, C. T., Baumann, S., & Grice, M. (2022). The influence of expectations on tonal cues to prominence. *Journal of Phonetics, 94*, 101174. https://doi.org/10.1016/j.wocn.2022.101174.

Röhr, C. T., Grice, M., & Baumann, S. (2023). Intonational Preferences for Lexical Contrast and Verum Focus. In: Radek Skarnitzl & Jan Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences (pp. 1578-1582). Guarant International. In R. Skarnitz & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1578–1582). Guarant International.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*, 51–89. https://doi.org/10.1016/S0010-0277(03)00139-2.

Schafer, A. J., Carter, J., Clifton, C., & Frazier, L. (1996). Focus in relative clause construal. *Language and Cognitive Processes, 11*(1–2), 135–163. https://doi.org/10.1080/016909696387240.

Schafer, A. J., & Jun, S.-A. (2002). *Effects of accentual phrasing on adjective interpretation in Korean* (M. Nakayama, Ed.; pp. 223–255). CSU. /paper/Effects-of-accentual-phrasing-on-adjective-in-Schafer-Jun/941d1876e59c6de8250332840cba9c311058e901.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193–247. https://doi.org/10.1007/Bf01708572.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language, 48*(1), 103–130. https://doi.org/10.1016/S0749-596X(02)00519-3.

Steffman, J. (2021). Prosodic prominence effects in the processing of spectral cues. *Language, Cognition and Neuroscience, 36*(5), 586–611.

Steffman, J., Kim, S., Cho, T., & Jun, S.-A. (2022). Prosodic phrasing mediates listeners' perception of temporal cues: Evidence from the Korean Accentual Phrase. *Journal of Phonetics, 95*, 101156.

Steffman, J., & Sundara, M. (2023). Disentangling the role of biphone probability from neighborhood density in the perception of nonwords. *Language and Speech*, 00238309231164982. https://doi.org/10.1177/00238309231164982.

Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience, 2*(2), 191–196. https://doi.org/10.1038/5757.

Toscano, J. C., & McMurray, B. (2015). The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments. *Language, Cognition and Neuroscience, 30*(5), 529–543. https://doi.org/10.1080/23273798.2014.946427.

Turco, G., Braun, B., & Dimroth, C. (2014). When contrasting polarity, the Dutch use particles, Germans intonation. *Journal of Pragmatics, 62*, 94–106. https://doi.org/10.1016/j.pragma.2013.09.020.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. *Language and Speech, 49*(3), 367–392. https://doi.org/10.1177/00238309060490030301.

Zahner-Ritter, K., Chen, Y., Dehé, N., & Braun, B. (2022). The prosodic marking of rhetorical questions in Standard Chinese. *Journal of Phonetics, 95*, 101190. https://doi.org/10.1016/j.wocn.2022.101190.