Language and Cognition (2023), 15: 3, 453–479
doi:10.1017/langcog.2023.8

ARTICLE

Chunking up speech in real time: linguistic predictors and cognitive constraints

Svetlana Vetchinnikova[1]*, Alena Konina[2], Nitin Williams[2,3], Nina Mikušová[2] and Anna Mauranen[2]

[1]Helsinki Collegium for Advanced Studies, University of Helsinki, Helsinki, Finland; [2]Department of Languages, University of Helsinki, Helsinki, Finland; [3]Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland
*Corresponding author. Email: svetlana.vetchinnikova@helsinki.fi

(Received 17 June 2022; Revised 22 February 2023; Accepted 01 March 2023)

Abstract

There have been some suggestions in linguistics and cognitive science that humans process continuous speech by routinely chunking it up into smaller units. The nature of the process is open to debate, which is complicated by the apparent existence of two entirely different chunking processes, both of which seem to be warranted by the limitations of working memory. To overcome them, humans seem to both combine items into larger units for future retrieval (usage-based chunking), and partition incoming streams into temporal groups (perceptual chunking). To determine linguistic properties and cognitive constraints of perceptual chunking, most previous research has employed short-constructed stimuli modeled on written language. In contrast, we presented linguistically naïve listeners with excerpts of natural speech from corpora and collected their intuitive perceptions of chunk boundaries. We then used mixed-effects logistic regression models to find out to what extent pauses, prosody, syntax, chunk duration, and surprisal predict chunk boundary perception. The results showed that all cues were important, suggesting cue degeneracy, but with substantial variation across listeners and speech excerpts. Chunk duration had a strong effect, supporting the cognitive constraint hypothesis. The direction of the surprisal effect supported the distinction between perceptual and usage-based chunking.

Keywords: chunking; speech perception; neural oscillations; prosody; syntax; surprisal; individual variation; cue degeneracy; syntagmatic redundancy; linguistic robustness

1. Introduction

The term chunking has a long history originating from Miller's (1956) famous paper "The Magical Number Seven, Plus or Minus Two". It most strongly associates with recoding sequential information into meaningful chunks to overcome the limitations of short-term memory. The example of an individual, SF, who was able to remember sequences of 79 digits by recoding them into running times and dates is well known (Ericsson et al., 1980). This principle seems to apply to language exceptionally well since in essence, it comprises a large inventory of chunks at different levels of organization from phonology to discourse (Christiansen & Chater, 2016; Ellis, 2017; Goldberg, 2003). As a result, there is a strong belief among language scientists that we process speech by chunking it up into multi-word units of some kind, ranging from lexically specified combinations to more abstract constructions. However, multi-word units are language chunks that we have learned from previous language experience, just as running times and dates SF used. What is often overlooked is that in memorizing a telephone number, it also helps to simply group the digits into strings of three or four even without further recoding to running times or the like (Hitch et al., 1996; Ryan, 1969; Wickelgren, 1964). Thus, it appears that there are two different chunking processes operating at the same time: on the one hand, we draw on the inventory of chunks available to us (usage-based chunking) and on the other, we segment incoming stream into temporal groups (perceptual chunking). While usage-based chunks are units of meaning and memory (Ellis, 2017), perceptual chunks are units of real-time processing (Sinclair & Mauranen, 2006).

Following Terrace (2001), Gilbert et al. (2015) draw a similar distinction between domain-general input chunking determined by the capacity of short-term memory and output chunking involving learned units stored in long-term memory. They point out that there is ample evidence from other domains that humans and animals produce and perceive continuous sequences in temporal groups marked by final lengthening. However, in contrast to sequences of digits and nonsense syllables, language has structure that has evolved as a result of cognitive, linguistic, and social constraints. Being a cognitive mechanism, perceptual chunking must have left a trace too: there are likely to be linguistic properties that signal perceptual chunk boundaries. What are they?

Given that in linguistics chunking is mostly viewed as recoding (e.g. Christiansen & Chater, 2016), perceptual chunking has mostly been studied in neuroscience. In fact, perceptual chunk boundaries seem to associate with a distinct event-related potential (ERP), the closure positive shift (CPS). Originally it was observed in relation to prosodic boundaries (Bögels et al., 2011; Steinhauer et al., 1999 for a review). However, it later became apparent that prosodic cues are neither a necessary nor a sufficient condition for a CPS to occur. For example, 3-year-olds did not show a CPS in response to prosodic cues unless there was a pause too (Männel et al., 2013). In adults, the CPS was observed in relation to commas during silent reading and simply after long constituents (Drury et al., 2016; Hwang & Steinhauer, 2011). Syntax interacted with prosody and modulated the amplitude of the CPS (Itzhak et al., 2010; Kerkhofs et al., 2007). Also, syntactically predictable phrase boundaries elicited a CPS in the absence of prosodic cues (Itzhak et al., 2010). Finally, when participants read identical three-clause sentences at different presentation rates, a CPS was observed at 2.7 s irrespective of the number of clauses that fit that time window, suggesting that perceptual chunking may be time-driven (Roll et al., 2012).

Time is also crucial in oscillation-based models of speech segmentation. Recent research shows that periodic neural oscillations at different frequency bands may be involved in “packaging incoming information into units of the appropriate temporal granularity” (Giraud & Poeppel, 2012, p. 511). By aligning with the speech dynamics at different timescales, oscillatory activity can set the temporal window for decoding and making predictions. It is fairly established that oscillations in the theta frequency band (4–8 Hz) entrain to or synchronize with, according to different accounts, the syllabic rhythm of speech, which is remarkably stable within and across languages

https://doi.org/10.1017/langcog.2023.8 Published online by Cambridge University Press

(Ding et al., 2017; Varnet et al., 2017). This phase synchronization to syllabic rate is also a prerequisite for intelligibility (Doelling et al., 2014; Ghitza & Greenberg, 2009). Similarly, when chunks of digits are presented at a fast rate outside the delta-band defined as 0.5–2 Hz, both oscillatory tracking and task performance are impaired (Rimmele et al., 2021). Thus, both the time window of delta oscillations and the memory constraint of 2–3 s, which roughly correspond to each other, suggest the existence of an optimal chunk duration. However, the relevant segmentation unit at this timescale is unclear.

One obvious candidate for a relevant segmentation unit is an intonation unit that seems to form a consistent rhythm at approximately 1 Hz across languages (Auer, 1999; Inbar et al., 2020; Stehwien & Meyer, 2021). However, several studies have observed oscillatory tracking of syntactic structure while controlling for prosody. Ding et al. (2016) found tracking of phrases at 2 Hz and clauses at 1 Hz in synthetically generated isochronous stimuli such as *dry fur rubs skin* where each word consisted of one syllable, each phrase of two words, and each clause of two phrases. Kaufeld et al. (2020) used more naturalistic stimuli such as *[Timid heroes] [pluck flowers] and the [brown birds] [gather branches]* and found tracking at the phrasal (0.8–1.1 Hz) and lexical (1.9–2.8 Hz) timescales (cf. Keitel et al., 2018). Finally, Henke and Meyer (2021) found that delta oscillations themselves can enforce segmentation at 2.7 s. They also suggest that the CPS discussed above is the time-domain equivalent of a delta-band phase reset (Meyer et al., 2017, 2020).

Thus, it appears that perceptual chunking may be driven by acoustic–prosodic cues, syntactic structure, or simply the optimal chunk duration reflecting the memory constraint of 2–3 s and/or the frequency of delta-band oscillations. Disentangling different linguistic cues and cognitive constraints is difficult especially since the short, constructed stimuli commonly used in previous research do not allow to examine more than one or two cues at a time. In addition, experiment results based on constructed sentences may not generalize to processing natural speech. For example, in contrast to writing, speech includes a large proportion of non-clausal material (NCM) and often defies a strict separation of syntax from prosody. In this study, we attempt to overcome these problems by working with natural language data. We extracted short excerpts of speech from linguistic corpora and played them to linguistically naïve listeners. While listening to the extracts, they intuitively marked chunk boundaries in the accompanying transcripts through a custom-built tablet application. In earlier research, we validated the method and showed that silent pauses inserted at intuitively marked chunk boundaries elicit a CPS while when inserted within a chunk they elicit a biphasic emitted potential suggesting interrupted processing (Anurova et al., 2022; Vetchinnikova et al., 2022). In this paper, we explored to what extent naïve listeners are affected by a range of different linguistic cues and cognitive constraints in chunk boundary perception. Specifically, we ask: (1) Which linguistic properties have an effect on chunk boundary perception and to what extent their effects vary across listeners and different speech samples? (2) To what extent does chunk duration constrain chunking? (3) Is there evidence for the dissociation between usage-based chunking and perceptual chunking?

Given the findings of previous research, we selected the following variables: syntax, prosody, pause, chunk duration, and bigram surprisal of the words before and after a chunk boundary. In Section 2, we discuss the main linguistic properties associated with chunking: syntax, prosody, and statistical regularities. In Section 3, we provide information about the participants, materials, and data collection procedures, explain how each of the variables was operationalized, and outline the statistical analysis. The results are presented in Section 4 and discussed in Section 5. Section 6 gives our conclusions.

2. Linguistic properties associated with chunking: clauses, prosodic units, or multi-word units?

2.1. Syntax

Which hierarchical level of syntactic constituent structure is relevant for perceptual chunking? Most grammars adopt the clause/sentence as the maximal unit of analysis. The clause is considered "the core unit of grammar" (Carter & McCarthy, 2006, p. 486) and is posited as the carrier of a message, "a quantum of information" (Halliday & Matthiessen, 2004, p. 58). Besides generative grammars, all major reference grammars for English (Biber et al., 1999; Carter & McCarthy, 2006; Huddleston & Pullum, 2002; Quirk et al., 1985) and influential functional grammars (Dik, 1997; Halliday & Matthiessen, 2004) similarly use the notion of a clause. Yet, the notion is inherited from the analysis of written language, since authentic spoken language data was not available to grammarians in large quantities until comparatively recently. It is now widely acknowledged that speech, in contrast to writing, does not consist of sentences (Biber et al., 1999; Carter & McCarthy, 1995; Leech, 2000).

Biber et al. (1999), a fully corpus-based reference grammar that includes a description of spoken English, point out that grammatical structure seems to be less important in speech compared to writing (cf. Halliday, 2009; Leech, 2000). They argue that while it is possible to analyze a stretch of spoken language in terms of embedding and coordination as in (1), this does not seem to be necessary, since the same stretch can be divided into a linear sequence of clause-like units, represented by vertical lines in (2): a mechanism they call the add-on strategy.

(1) [The trouble is [[if you're the only one in the house] he follows you] [and you're looking for him] [so you can't find him.]]]

(2) The trouble is | if you're the only one in the house | he follows you | and you're looking for him | so you can't find him.
adapted from Biber et al. (1999, p. 1068)

Despite this observation, Biber et al. (1999, pp. 1069–1070) adopt a larger unit of analysis, what they call a C-unit, defined as the maximal syntactically independent unit. C-units can be clausal and non-clausal. Clausal units include the main clause (MC) and all dependent clauses embedded within it. In other words, the entire utterance in (1) and (2) is one C-unit. Non-clausal units are segments that cannot be treated as part of any clausal units: according to their analysis, these account for 38.6% of C-units in conversational data.

In our syntactic annotation, we start from the assumption that we do not know which syntactic information listeners use for chunking natural speech and attempt to capture as much information about syntactic structure as possible. We draw on Biber et al.'s analysis but do not discard embedding.

2.2. Prosody

Many emphasize that the boundaries of prosodic and syntactic units tend to coincide. However, the nature and the extent of the correspondence remain a bone of contention (Cole, 2015; Frazier et al., 2004; Wagner & Watson, 2010; Watson & Gibson, 2004). In the literature, broad agreement prevails that prosodic boundaries are sensitive to a variety of factors, including syntactic. Research into prosody-syntax relationships has moved from predominantly theory-based structural modeling toward a more empirical foundation and growing attention to the complex interconnectedness of linguistic systems, processing factors, and contextuality. Early prosodic research drew heavily on generative traditions, emphasizing underlying structure in the phonological component analogically to syntax, even when describing a grammar of intonation without explicitly invoking syntax (Pierrehumbert, 1980). By contrast, some descriptions (such as Truckenbrodt's, 1999) assumed a tight interdependency between syntactic and prosodic phrasing, postulating a constraint (WRAP-XP) that demands each syntactic phrase to be contained in a phonological phrase. Selkirk (1978), in turn, held that prosodic structure is separate from syntax and not isomorphic to it, but nevertheless argued for an interrelationship that enables mapping prosodic structure onto generative syntax. Moreover, she stressed the connectedness of prosody and meaning, articulated in her Sense Unit Condition (Selkirk, 1984), which stipulates that constituents of an intonational phrase must form a sense unit together. Later experiments involving naturalness judgments on generalized versions of Truckenbrodt's Wrap and Selkirk's Sense Unit Condition (Frazier et al., 2004) found support only for the latter.

Ferreira (1993, 2007) incorporated semantic constraints into predictions of segmental properties, which resonates with Selkirk's (1984) notions. She also suggested a trading relationship between pause duration and pre-pausal lengthening (Ferreira, 1993). Beyond such local interactions between different system components, prosodic elements are also implicated beyond sentence boundaries. Gee and Grosjean (1984) investigated pausing in narratives and concluded that pausing is sensitive not only to sentence-level prosody but also to narrative structure.

Overall, prosody research has been moving from theory-based arguments and descriptions toward varied experimental designs and contexts. Hypotheses pinned on decontextualized, read-aloud written sentences have given way to naturalness judgments (e.g. Frazier et al., 2004) and relatively natural contexts such as cooperative game tasks (Schafer et al., 2000) where speakers' prosodic phrasings show more variability. As Schafer et al. note, readers and conversationalists have dramatically different pragmatic goals. Moreover, Wagner and Watson (2010) observe that since there is more than one way of constructing complex meanings and presenting the same syntax prosodically, speakers' choices may ultimately lie in processing factors.

Moving toward more naturalistic data has made researchers increasingly aware that syntax and prosody are inevitably immersed in context. Choices are less invariant than has been assumed, as illustrated by Frazier et al.'s (2004) observation that listeners are concerned with relative, not absolute break size in local contexts. Human languages thus seem to have flexible systems allowing appreciable optionality in prosodic phrasing of syntactic units such as sentences (Schafer et al., 2000). Cole (2015) goes further, suggesting that prosody research should embrace contexts of various kinds. She strongly advocates methods eliciting interactive speech from speakers engaged in genuine communication with meaningful discourse goals. This is also highly relevant for developing spoken dialogue systems for human-machine interaction in computational linguistics (e.g. Edlund & Heldner, 2005; Swerts & Hirschberg, 1998, 2008).

https://doi.org/10.1017/langcog.2023.8 Published online by Cambridge University Press

Altogether, prosody is flexible and complex, as is syntax, affected by processing and embedded in context, and clearly, it is both desirable and possible to tackle them in spontaneous speech.

2.3. Surprisal

It is known that comprehenders are sensitive to statistical regularities in the input and use them to discern structure. Saffran et al. (1996), a classic study in statistical learning, showed that even 8-month-old infants can use transitional probabilities between syllables to break continuous input into word-like units. McCauley and Christiansen (2014, 2019) developed a computational model, the chunk-based learner (CBL), to show that the same mechanisms can account for chunking input into larger, multi-word units.

Statistical learning studies have traditionally focused on transitional probabilities between sequential elements of the input, which are higher within recurring units and lower across them, causing dips in transitional probability associated with unit boundaries. Forward and backward transitional probabilities are possible alternatives: the former are more in line with the view of language comprehension as a predictive process, and the latter work better for cases such as *the dog* where *the* is much more likely to precede *dog* than *dog* to follow *the,* since *the* can be followed by any noun. Studies of natural language comprehension often adopt surprisal as a measure of linguistic prediction which calculates how predictable a word is given its context, such as the previous word (Levy, 2008; Shain et al., 2020). To illustrate, *sort* is almost always followed by *of*, so *of* will be expected by the listener and its surprisal will be low. Predictable words are processed faster (Smith & Levy, 2013) and have a smaller N400 amplitude (Frank et al., 2015).

In this study, we hypothesize that if perceptual chunks are learned multi-word units, the word before a boundary should be more predictable, while the word after a boundary should be less predictable. To examine the effect of statistical information on chunk boundary perception, we will include bigram surprisal for the word preceding a boundary (*closing surprisal*) and the word following it (*opening surprisal*).

3. Methods

All materials, data, and code are openly available in the Open Science Framework repository at https://osf.io/7bta5/.

3.1. Participants

We recruited 51 volunteer students from different disciplines of the University of Helsinki excluding language sciences. They were fluent nonnative speakers of English with a variety of first language backgrounds, aged 20–39 (36 females, 44 right-handed). None reported dyslexia. All volunteers submitted informed consent before the experiment and received a movie ticket for their participation. One participant's data were discarded from the final dataset due to clear inactivity during the main experiment task (marked only 12 boundaries in total).

3.2. Materials

As discussed in the introduction, naturally occurring speech is very different from writing. One of the main aims of the experiment was to study chunking in authentic speech. For this reason, the speech stimuli were extracted from three corpora of natural, native, and nonnative English speech recorded in university environments: the Michigan Corpus of Academic Spoken English (MICASE), the Corpus of English as a Lingua Franca in Academic Settings (ELFA) and the Vienna-Oxford International Corpus of English (VOICE). Typical speech events represented in the corpora include lectures, seminars, conference presentations, and discussions. Using automatic and manual means, we identified 97 10-to-45-s-long excerpts (M = 55 words, SD = 14, min = 29, max = 100, and total = 5237 words) which were fluent and intelligible without wider context. We avoided unintelligible or unfinished words, laughter, long pauses, overlapping speech, speaker changes, frequent hesitations, or repetitions and controlled for specialized and low-frequency vocabulary: some of these criteria are stipulated by a parallel brain imaging experiment we report in Anurova et al. (2022). Since the audio quality of the original extracts was uneven, we recruited a speaker who read out the extracts mimicking the prosodic patterns of the original audio clips as close as possible. The recordings were made in an acoustically shielded studio at the phonetics laboratory of the University of Helsinki.

3.3. Procedure

In the experiment, each participant received a tablet and headphones and was asked to follow the instructions on the screen. The workflow contained a consent form, a background questionnaire, the chunking task itself, an Elicited Imitation task as a quick language proficiency test (Culbertson et al., 2020), and a feedback form. Personal information was not collected.

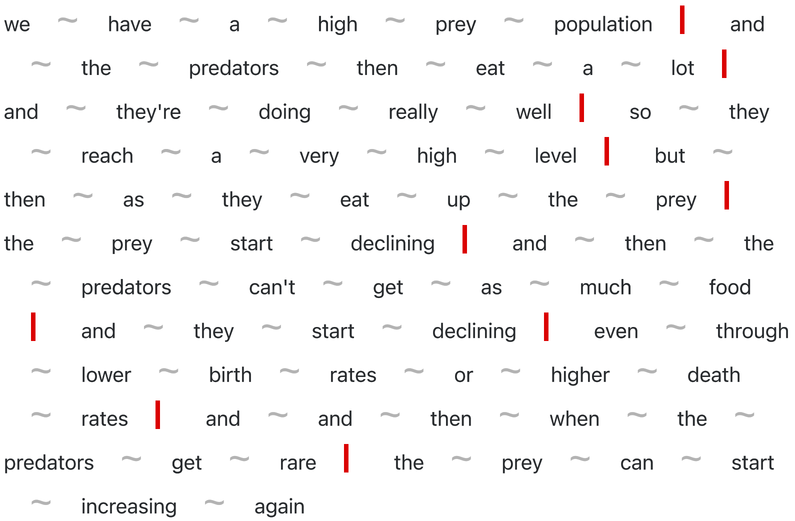
The chunking task was performed via a custom web-based tablet application *ChunkitApp* (Vetchinnikova et al., 2017, 2022; cf. Cole et al., 2017; https://www.chunkitapp.online/). Participants listened to the audio clips and simultaneously marked chunk boundaries in the transcripts displayed on the screen (see Appendix for full instructions). The notion of a chunk was not explained, encouraging participants to work intuitively. All orthographic words in the transcripts were separated with a tilde symbol (~) which one could tap to insert or remove a boundary (see Fig. 1). Each audio clip was played once only. While an aural-only presentation would approximate natural speech comprehension better, it is difficult to implement this technically since listeners may react slightly earlier, in anticipation of the coming boundary, or slightly later, as a post hoc realization of the past boundary obscuring the exact locations of chunk boundaries.

Each audio clip was followed by a self-evaluation (75%) or a true/false comprehension question (25%) to keep the attention of the participants on the task and to probe comprehension. The whole experiment session took up to 2 h including a coffee break. The participants could take additional breaks at any time.

3.4. Predictor variables

For convenience, in what follows we will use the term *word* for any string of letters separated by spaces in the transcripts of the stimulus extracts since each space could

https://doi.org/10.1017/langcog.2023.8 Published online by Cambridge University Press



we ~ have ~ a ~ high ~ prey ~ population | and
~ the ~ predators ~ then ~ eat ~ a ~ lot |
and ~ they're ~ doing ~ really ~ well | so ~ they
~ reach ~ a ~ very ~ high ~ level | but ~
then ~ as ~ they ~ eat ~ up ~ the ~ prey |
the ~ prey ~ start ~ declining | and ~ then ~ the
~ predators ~ can't ~ get ~ as ~ much ~ food
| and ~ they ~ start ~ declining | even ~ through
~ lower ~ birth ~ rates ~ or ~ higher ~ death
~ rates | and ~ and ~ then ~ when ~ the ~
predators ~ get ~ rare | the ~ prey ~ can ~ start
~ increasing ~ again

Fig. 1. User interface, *ChunkitApp.*

be marked as a chunk boundary in the experiment. In other words, each word can be potentially followed by a chunk boundary.

3.4.1. Syntactic boundary strength

To make syntactic annotation maximally informative, we applied a traditional hierarchical analysis of the constituent structure and focused on identifying clauses as the most likely syntactic analog of a chunk. Since the aim of the study was to disentangle different cues and find out how much they contribute to chunk boundary perception, we limited the syntactic annotation to structural information, ignoring semantic, pragmatic, or prosodic information as far as possible, by for instance working from transcripts without consulting the audio recordings.

We defined a clause as a constituent structured around a verb phrase, including both finite and nonfinite clauses. Dependent clauses were allowed to be embedded within the MCs. All clauses were identified and tagged (see Example 3). Material that fell outside the constituent structure of a clause was annotated as non-clausal. Examples of NCM include hesitations (*er*, *erm*, *uh*), repetitions, rephrases, pragmatic markers (*all in all*, *of course*, *basically*, *sort of*), and unembedded dependent clauses. In contrast, units such as *I mean*, *I thought*, *as we all know* were analyzed as clausal even though pragmatically they are likely to function as discourse markers.

Example 3 shows the annotation of an extract. Clauses are marked with square brackets, NCM with round brackets. Since the analysis is hierarchical and multiple embedding is allowed, boundaries where more than one clause start or finish are marked with multiple square brackets as in line 3, where a non-finite infinitive clause (NF-to) is embedded within an MC, and line 6, where an NF-to is embedded within a relative clause (DC-R) which is itself embedded in an MC.

https://doi.org/10.1017/langcog.2023.8 Published online by Cambridge University Press

(3)

Line	Extract	Tags	Value
1	(about the methods)	NCM	0.5
2	[one choice would be		0.5
3	[to study this issue only theoretically]]	NF-to MC	2.5
4	[but at the moment I prefer another choice		0.5
5	[which would be		0.5
6	[to include to this study an empirical part]]]	NF-to DC-R MC	3

Following this annotation, syntactic boundary strength can be operationalized in several different ways: (a) as a categorical variable with four levels distinguishing between non-clausal/non-clausal, clausal/non-clausal, non-clausal/clausal, and clausal/clausal boundaries, (b) as a categorical/continuous variable based on the number of clausal boundaries (square brackets) resulting in seven categories (0–6), and (c) a categorical/continuous variable based on the weighted number of clausal boundaries to distinguish between opening and closing brackets and resulting in ten categories (0–5.5, with no boundaries with exactly 4 or 5 clausal boundaries). The last operationalization assumes that in chunk boundary perception the end of a clause is more important than the start of a new one and assigns 0.5 for each opening bracket and 1 for each closing bracket. Thus, for example, the clausal boundary at the end of line 3 in Example 3 is assigned the value 2.5 since there are two closing brackets and one opening bracket.

To test these assumptions and select the maximally informative operationalization, we compared the operationalizations and their relationship with boundary marking using chi-square tests. Table 1 gives the results of the chi-squared tests.

Syntactic boundary strength operationalized as the weighted number of clausal boundaries returned the largest effect size suggesting that both the number of ending/starting clauses after a given word and the distinction between ending and starting clauses contribute to the effect. The weighted number of clausal boundaries was selected as the maximally informative operationalization of syntactic boundary strength.

3.4.2. Prosodic boundary strength

Prosodic boundary strength was estimated with the Wavelet Prosody Toolkit (Suni, 2017; Suni et al., 2017), a computer program that calculates predicted prosodic boundary strength and prominence estimates for speech signals in an unsupervised fashion. The program aligns the speech signal with the transcript, extracts prosodic signals of the fundamental frequency, energy, and word duration (excluding pauses

Table 1. Results of the chi-squared tests relating boundary markings to different operationalizations of syntactic boundary strength

Syntactic boundary strength	χ^2	Two-tailed *p*	Cramer's *V*	Bootstrapped 95% CI	df
Four-way classification	81237	*<0.001*	0.562	[0.557; 0.568]	3
No. of clauses	78514	*<0.001*	0.553	[0.547; 0.558]	6
Weighted no. of clauses	83118	*<0.001*	0.569	[0.564; 0.575]	9

and breaths), and combines them. Then it applies the continuous wavelet transform (CWT) to decompose the composite signal into scales that roughly correspond to the levels of prosodic hierarchy: syllables, words, and phrases. The method assumes that both word prominences and prosodic boundaries arise from the same sources of the signal. Thus, the peaks formed by the signal at different scales indicate prominences while the troughs indicate boundaries. Hierarchically organized peaks across the scales are joined into a line of maximum amplitude expressing prominence strength and troughs into a line of minimum amplitude expressing boundary strength. The program produces continuous values of prominence and boundary strength for each word. In our data, prosodic boundary strength varies between 0 and 2.436.

The method was evaluated on the manually ToBI annotated Boston Radio News corpus (Ostendorf et al., 2005) and showed 84.6% accuracy for prominence detection and 85.7% accuracy for boundary detection outperforming other unsupervised methods (Suni et al., 2017). Thus, the method closely approximates human processing of speech prosody but being unsupervised and purely signal based, avoids the problems associated with human annotators, such as subjectivity, variability, and possible influence of other linguistic cues.

3.4.3. Pause duration

Orthographic and phonetic alignment of the transcripts to their audio files was completed automatically with WebMAUS (Schiel, 1999) and then manually corrected in Boersma & Weenink (2022). Pause duration is the time between the end of any given word and the start of the next word.

3.4.4. Chunk duration

The temporal constraint hypothesis rooted in the limitations of working memory capacity and/or the periodicity of neural oscillations predicts that chunks should be limited as well as fairly regular in their duration. In other words, increasing duration should associate with a higher likelihood of chunk boundary perception.

The procedure of calculating chunk duration for each word in the extracts starts with the identification of chunk onsets. In Vetchinnikova et al. (2022), we proposed that chunks can be identified via crowdsourcing chunk boundary perception data and finding those places where inter-rater agreement on chunk boundary is statistically significant. To test inter-rater agreement rates for statistical significance, we used permutation tests. Individual boundary markings (that is zeros and ones for each individual) were permuted one million times (with replacement) to obtain the null distribution for each boundary. Two-tailed p-values were calculated by comparing each observed boundary frequency with one million permuted ones and finding how many times the observed or a more extreme boundary frequency occurs in the permutations. To avoid zero p-values in cases where the observed or a more extreme boundary frequency did not occur in the permutations, we defined p as the upper bound $p_u = (b + 1)/(m + 1)$ where b is the number of times permuted boundary frequency is equal or more extreme than the observed and m is the number of permutations (Phipson & Smyth, 2010; Puoliväli et al., 2020). The p-values were corrected for false discovery rate (FDR) at $\alpha = 0.05$ (Benjamini & Hochberg, 1995; Puoliväli et al., 2020). Boundaries with boundary frequencies lower than expected by chance (≤ 0) were considered statistically significant non-boundaries and boundaries with boundary frequencies higher than expected by chance (≥10) were statistically significant boundaries. Thus, in this dataset chunks are defined as strings of words between boundaries marked by 10 and more listeners.

The resulting chunks proved to be fairly regular in their duration ($M = 2.55$ s, SD = 1.2 s) which already partly answers the research question. However, the average duration does not show the extent of individual variability in the effect of duration on chunk boundary perception since it is based on aggregate boundary markings. Possible variability across extracts is also not taken into account. Thus, using identified chunk boundaries, we calculated chunk duration for each word in order to include it in the model predicting boundary markings at the individual level.

For any given word, chunk duration is calculated from the onset of a chunk to the onset of the next word. For example, in the chunk *we have a high prey population* (Table 2), the duration for the word *prey* is 1.04 s, which is calculated from the onset of the word *we* until the onset of the word *population*. The running time is reset at the onset of the word *and* since a statistically significant number of listeners (32, $p < 0.05$) mark a boundary after the word *population*.

Since listener agreement on chunk boundaries is used in the identification of chunk onsets, the operationalization of chunk duration may be to a certain extent confounded with agreement. We will address this possibility in Section 4.2.

3.4.5. Surprisal

To estimate surprisal for each word in our stimuli, we compiled a separate reference corpus of academic speech since all existing general reference corpora are either too biased toward written language (the Corpus of Contemporary American English [COCA] or the British National Corpus [BNC]) or sampled from a range of registers and language varieties incomparable with ours (e.g. the spoken component of the BNC2014). Our purpose-built reference corpus contains the three corpora the experiment stimuli were selected from: MICASE, ELFA, and VOICE. To increase size, we also added the British Academic Spoken English (BASE) corpus, which was designed as a companion to MICASE and therefore fully comparable, and a corpus of TED talks subtitles (Reimers & Gurevych, 2020; Tiedemann, 2012), which was deemed close enough to academic lectures. The resulting corpus has 12.5 million words and contains 670 text files (Table 3).

For each bigram AB, surprisal of B was calculated as −log2(conditional probability of B), where the conditional probability of B was estimated as corpus frequency of AB divided by corpus frequency of A. Bigram surprisal was calculated for each word in the dataset. We examined the effect of word surprisal on the probability that a

Table 2. Example of annotation for chunk duration

Word	*we*	*have*	*a*	*high*	*prey*	*population*	*and*	*the*	*predators*	*then*	*eat*	*a*	*lot*
Boundary frequency	0	0	0	1	1	**32**	0	0	0	0	1	0	29
Duration	0.09	0.28	0.35	0.74	1.04	2.27	**0.08**	0.17	0.72	1.15	1.63	1.70	2.39

https://doi.org/10.1017/langcog.2023.8 Published online by Cambridge University Press

Table 3. A reference corpus of academic speech compiled to estimate surprisal

Sub-corpus	No. of words	No. of files
ELFA	1,013,666	167
MICASE	1,666,539	152
VOICE 2.0	991,336	151
BASE	1,635,606	199
TED2020 v1	7,202,498	4076
Total	12,509,645	4745

chunk boundary would follow it (*closing surprisal*, A**B**~) or precede it (*opening surprisal*, A ~ **B**).

3.5. Statistical analysis

All analyses were conducted in R version 4.2.2 (R Core Team, 2022). Our predictor variables were pause duration, prosody or prosodic boundary strength, syntax or syntactic boundary strength (operationalized as a weighted number of clausal boundaries), chunk duration, closing surprisal, and opening surprisal. All variables except closing surprisal and opening surprisal are positively skewed with a large number of zero values (Fig. 2). This is unsurprising since we can expect only one chunk boundary per 5–10 words and it is reasonable to assume that the predictor variables are distributed in a similar way. The response variable is binary: each participant could either mark a boundary (1) or leave it unmarked (0). The data points are non-independent and required the inclusion of random intercepts and slopes by listener and by extract. We expected collinearity between the variables: strong chunk boundaries are likely to occur at the end of a clause and be marked by

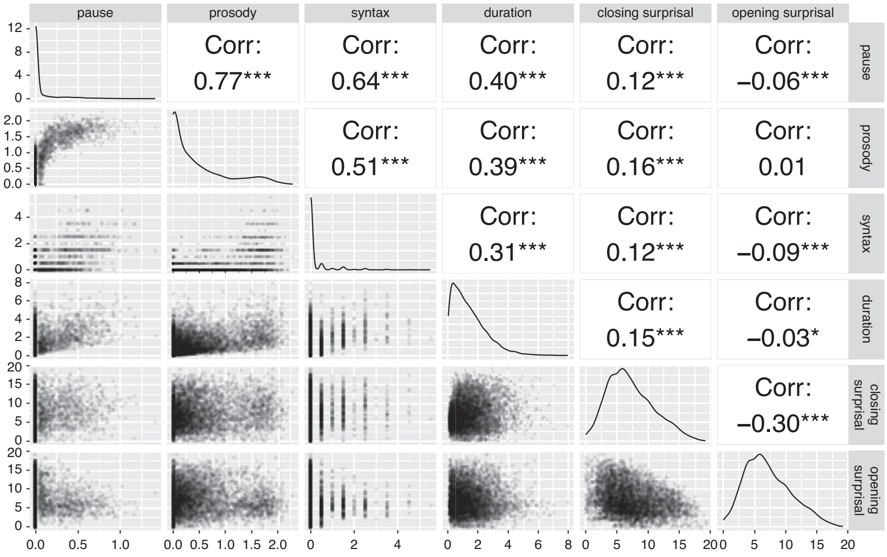


Fig. 2. Scatter plot matrix for all independent variables indicates collinearity.

https://doi.org/10.1017/langcog.2023.8 Published online by Cambridge University Press

both prosody and a longer pause. The scatter plot matrix in Fig. 2 shows that pause, prosody, and syntax are indeed positively intercorrelated.

Pause and prosody show the strongest relationship ($r = 0.77$). The relationship between syntax and prosody is less pronounced ($r = 0.51$), giving support to those linguistic theories which question full alignment between the two properties. Chunk duration is moderately correlated with all three linguistic variables expected to predict chunk boundary perception ($r = 0.3–0.4$). Closing surprisal and opening surprisal do not correlate with any variables but have a small negative association with each other ($r = -0.3$). The scatterplot shows that the bulk of the points falls into the lower left quadrant, with the upper right quadrant virtually empty, suggesting that while it is common for words with surprisal below the mean to follow each other, words with surprisal above the mean almost never do. Cases, where a word with low surprisal is followed by a word with high surprisal (lower right quadrant), are probably represented by combinations of function words with content words (Table 4, examples 7–12). Cases, where a word with high surprisal is followed by a word with low surprisal (upper left quadrant), include both combinations of content words with function words and multi-word units where the first word is a rare word while the following word is predictable based on the first word (Table 4, examples 1–6).

Due to collinearity between independent variables, we examined the effect of each predictor on chunk boundary perception in isolation to see how much variance each predictor can explain alone. For this, we used the *lme4* package version 1.1–29 (Bates et al., 2015) to fit a series of mixed logistic regression models, each of which estimates the probability of boundary marking according to the following syntax:

$$\text{response} \sim \text{predictor} + (1 + \text{predictor}|\text{listener}) + (1 + \text{predictor}|\text{extract})$$

Individuals can vary in how often they mark a boundary as well as in the extent to which they rely on particular predictors to perceive a boundary. The extracts in our database were sourced from speech events in English corpora around the globe (Section 3.2). Even though re-recorded with one speaker, the original speakers were all different and can be expected to vary in the syntactic and prosodic timing and structuring of their speech. Our earlier study shows that extracts indeed vary in the

Table 4. Examples of word combinations with high-low and low-high surprisal pattern

	Word	Surprisal	Word	Surprisal
1	*preservation-minded*	18.28	*conservation*	0.00
2	*shuts*	17.88	*down*	0.72
3	*sleepless*	17.57	*nights*	0.12
4	*veterinary*	17.57	*college*	3.25
5	*visiting*	17.37	*scholars*	5.10
6	*computationally*	16.40	*intensive*	3.70
7	*the*	1.80	*heads*	14.10
8	*the*	0.58	*whole*	7.68
9	*in*	0.33	*explaining*	14.43
10	*of*	0.06	*identity*	12.02
11	*of*	0.06	*made*	14.04
12	*to*	0.03	*understand*	7.39

degree of listener agreement on chunk boundaries (Vetchinnikova et al., 2022), but they can also vary in the relative importance of different cues.

All variables were *z*-scored. For ease of interpretation, the effects plots use the original scale of the variables since zero is a meaningful value in all of them (e.g., no pause or no clausal boundary), while the mean is less informative due to the skewed distributions. The results tables were generated with the package *sjPlot* (Lüdecke, 2022) which returns marginal and conditional R-squared statistics, based on Nakagawa et al. (2017). To test the significance of the effect of each variable, we conducted likelihood ratio tests comparing each model to a null model of identical structure and control parameters, but the predictor of interest removed. To examine the extent of variation in the effects of different cues across listeners and across speech samples, we plotted random slopes and examined their distributions. We also examined slope/slope correlation matrices, to see whether different listeners may prefer different cues. Clearly, different speech samples may also foreground different cues. Since the analysis revealed five outlier listeners who did not seem to react to any of the cues (see Figs. 3 and 4), their random slope estimates were not included in the slope/slope

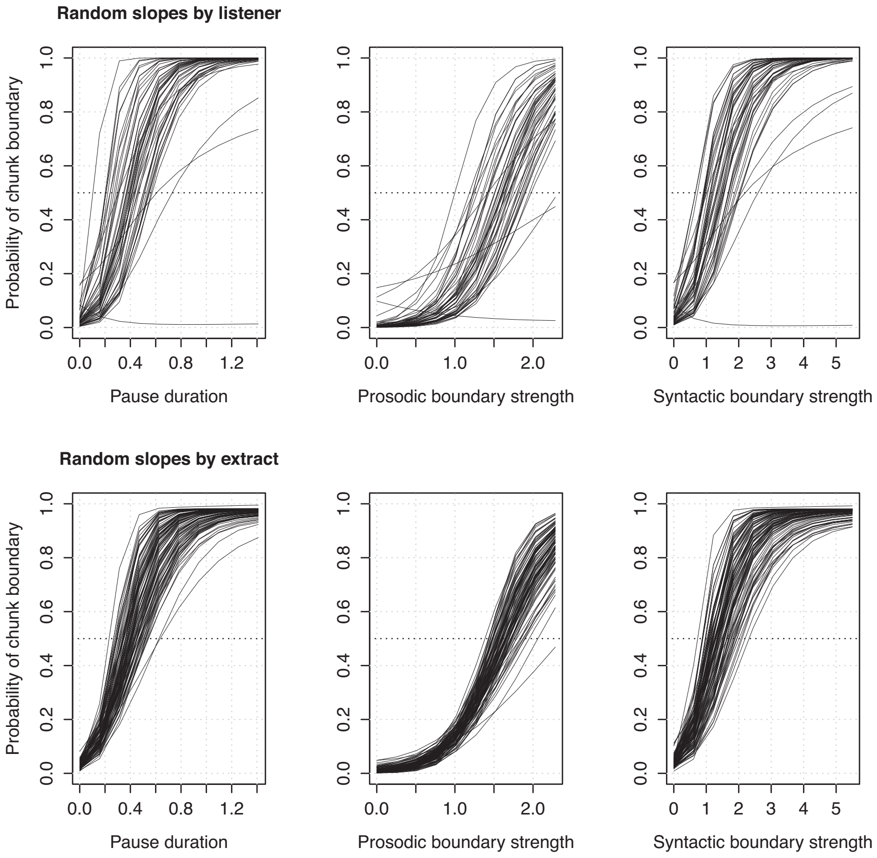


Fig. 3. Effects plots with by-listener and by-extract random slopes for pause, prosody and syntax.

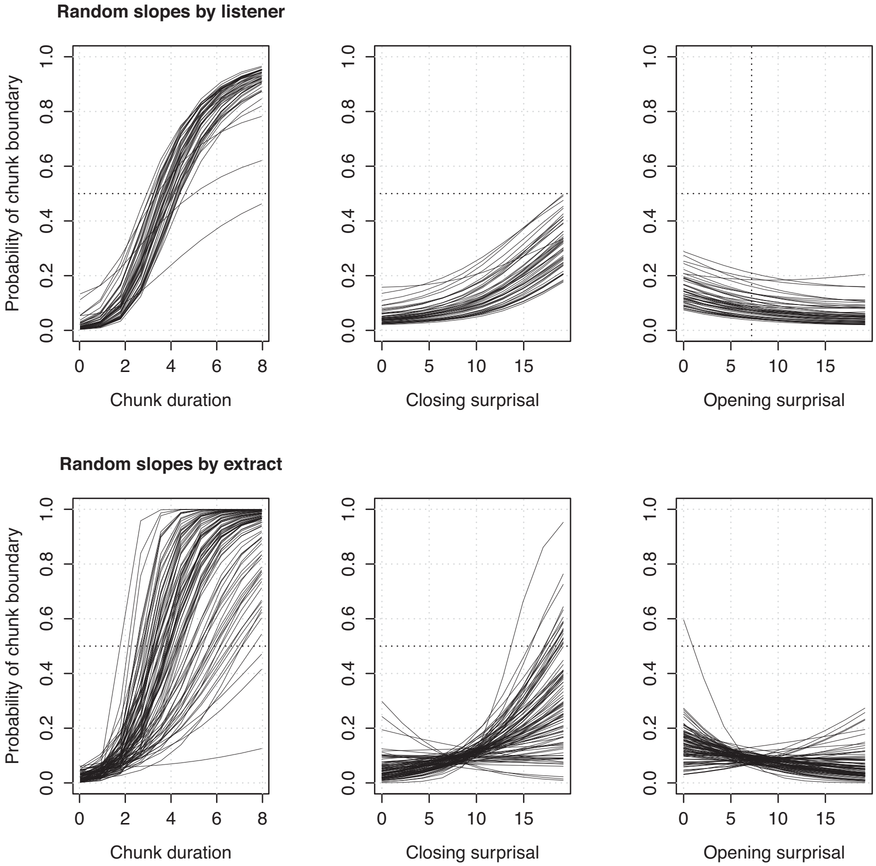


Fig. 4. Effects plots with by-listener and by-extract random slopes for duration, closing and opening surprisal.

correlation matrices to avoid their disproportionately large effect on correlation coefficients. This decision is further discussed in Section 5.

Based on the analysis of the single-predictor models, we also conducted some exploratory follow-up analysis described in Section 4.2.

4. Results

4.1. Single-predictor models

All single-predictor models show a significant effect of the predictor of interest (Table 5). As expected, pause, prosody, and syntax each have a positive effect, indicating that longer pauses and stronger prosodic and syntactic boundaries are more likely to be perceived as chunk boundaries. Chunk duration also has a positive effect, supporting the hypothesis that listeners are more likely to mark a boundary as chunks become longer. The effects of closing and opening surprisal are very small but interestingly the directions of the effects are opposite to statistical learning predictions: chunk boundaries associate with higher closing surprisal and lower opening

Table 5. Results of single-predictor models

Model	Pause	Prosody	Syntax	Duration	Closing surprisal	Opening surprisal
Fixed effects (log-odds)						
Intercept	−3.36	−4.05	−3.05	−2.96	−2.54	−2.46
95% CI	−3.62; −3.10	−4.33; −3.76	−3.29; −2.82	−3.18; −2.75	−2.69; −2.39	−2.60; −2.31
Slope	1.71	1.88	1.44	1.43	0.46	−0.26
95% CI	1.52; 1.91	1.72; 2.04	1.29; 1.59	1.27; 1.59	0.37; 0.54	−0.32; −0.19
Random effects						
Extract intercept variance	0.18	0.28	0.2	0.26	0.07	0.04
Listener intercept variance	0.8	0.91	0.59	0.46	0.25	0.24
Extract slope variance	0.16	0.14	0.14	0.48	0.17	0.09
Listener slope variance	0.4	0.26	0.21	0.06	0	0.01
Extract intercept/slope *r*	−0.29	−0.86	−0.09	0.12	−0.56	0.11
Listener intercept/slope *r*	0.16	−0.53	−0.01	−0.62	−0.14	0.5
ICC	0.32	0.33	0.26	0.28	0.13	0.1
Marginal R^2	0.378	0.420	0.317	0.310	0.053	0.018
Conditional R^2	0.576	0.609	0.494	0.502	0.175	0.120
Likelihood ratio test	103.67***	131.67***	112.91***	156.57***	75.75***	46.09***

***p< 0.001.

surprisal. In other words, the word before a boundary is less predictable while the word after a boundary is more predictable, in contradiction to the hypothesis that perceptual chunks are learned multi-word units.

In addition, random effects are important for all variables, as R^2 marginal is in all cases lower than R^2 conditional, which takes both fixed and random effects into account. All four major predictors can explain 50–60% of the variance alone. Prosodic boundary strength seems to be the strongest predictor accounting for the largest proportion of the total variance (60.9%), followed by pause duration (57.6%). Syntactic boundary strength and chunk duration account for about 50% each. However, these values are likely to be inflated since in separate models each predictor can absorb all the variability, especially as they are correlated.

Table 5 also shows the mean variances in the effects across listeners and extracts. In the following, Figs. 3 and 4 plot the effect of each variable on the predicted probability of perceived chunk boundary with by-listener and by-extract random slopes and intercepts. Fig. 5 shows random slope distributions and Fig. 6 shows slope/slope correlation matrices.

As shown in Figs. 3 and 4, the magnitude of each effect clearly varies both across extracts and across listeners, suggesting that listeners may be relying on different cues to different degrees, and that the reliability of cues in predicting chunk boundaries may differ by extract. Also, the listener sample clearly includes a few outliers who do not seem to react to any of the cues to the same extent as others, as their slopes are much flatter or even negative: these are the same people across the four major predictors. It is difficult to name the reasons for this divergent performance since these listeners answered the comprehension questions well enough and did not have the lowest scores on the proficiency test. However, they marked the largest number of boundaries which nobody else marked (one-off boundaries).

The effect of chunk duration stands out in Fig. 4 as it has much larger variability across extracts than across listeners. While listeners are consistently affected by chunk duration, extracts vary in the magnitude of the effect, which possibly reflects

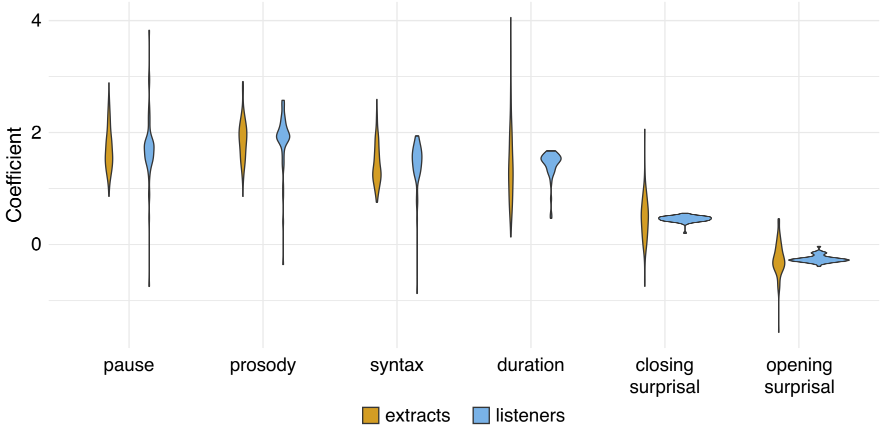


Fig. 5. Distribution of random effect slopes by extracts and by listeners.

different degrees of rhythmicity: if chunks are similar in duration, it should be easy to predict chunk boundaries simply based on timing. With regard to closing and opening surprisal, Fig. 4 shows how small the effects are, especially in comparison to other predictors.

The violin plots in Fig. 5 further highlight the variability of the effect of chunk duration across extracts. They also show that the effects of opening and closing surprisal are the only ones that vary in a direction across extracts, but not across listeners: while listeners are consistent in how they interpret surprisal, surprisal itself is not consistent in how it relates to chunk boundaries in different speech materials.

Scatter plots in Fig. 6 show that across the four major predictors, there is a strong correlation only between the by-listener effects of prosody and pause ($r = 0.75$), suggesting that those who rely on prosody also tend to rely on pause duration. At the same time, the correlations between the effects of syntax and prosody and syntax and pause are small to moderate ($r = 0.24$–0.38), suggesting that some listeners may have their individual preferences for the cues they track. The correlations between the effects of pause, prosody, and syntax by extract are also substantially smaller than the correlations between the variables themselves (Fig. 2), suggesting that even though the cues tend to converge, listeners may be tracking specific cues in different extracts.

Since the effects of opening and closing surprisal are small (Fig. 4), their relationships with other variables illustrated in Fig. 6 are mostly not statistically significant. However, there are a few puzzling by-listener effect correlations, such as a strong negative correlation between the effect of opening surprisal and chunk duration ($r = -0.71$). We reasoned that there might be a confounding factor: for example, there may be specific words with low surprisal which listeners associate with the opening of new chunks, such as conjunctions *and, but* and *so.* We tested this hypothesis in the exploratory analysis described in Section 4.2.

4.2. Exploratory analysis

As the analysis of the single-predictor models showed, although some relationship between word surprisal and chunk boundaries exists, it is not in line with the

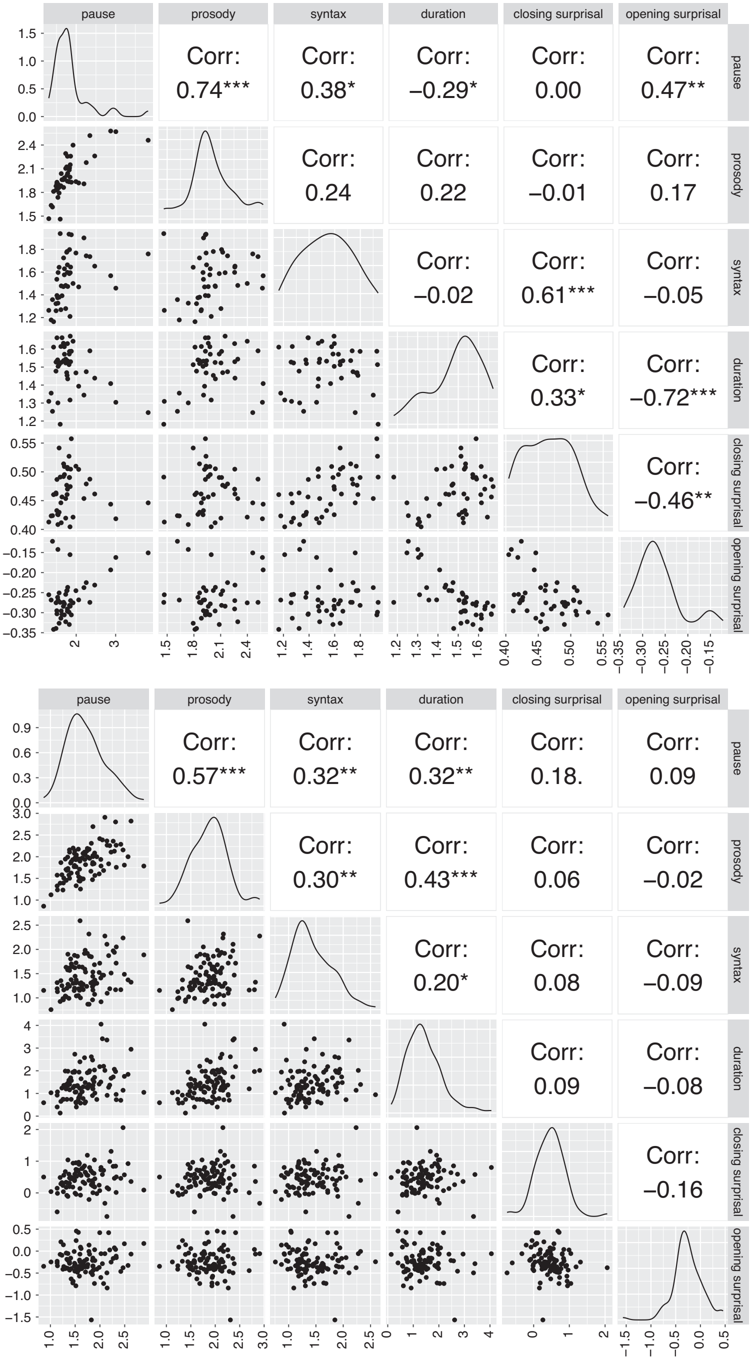


Fig. 6. By-listener (top) and by-extract (bottom) slope/slope correlations are smaller than expected, suggesting that listeners vary in cue preference and extracts vary in cue importance.

Table 6. Results of the single-predictor model based on whether the opening word is a conjunction

Model	Conjunction (yes)
Fixed effects (log-odds)	
Intercept	−2.8
95% CI	−2.98; −2.63
Slope	3.04
95% CI	2.65; 3.42
Random effects	
Extract intercept variance	0.1
Listener intercept variance	0.34
Extract slope variance	2.52
Listener slope variance	0.68
Extract intercept/slope correlation	−0.4
Listener intercept/slope correlation	−0.16
ICC	0.15
Marginal R^2	0.106
Conditional R^2	0.239

hypothesis that chunks are learned multi-word units. While more research is needed to understand the nature of the relationship, here we test whether some of the association between chunk boundaries and low opening surprisal can be explained by conjunctions *and, but* and *so*, which, as all function words, tend to be low in surprisal. We created a new variable conjunction with two levels: yes (the opening word is a conjunction) and no (the opening word is not a conjunction) and ran a mixed effects single-predictor model of the same type as in Section 4.1. The model was able to explain 24% of variance (Table 6). Further, when we added opening surprisal to this model, it was no longer a significant predictor suggesting that most of its effect can be explained by low surprisal of conjunctions which are associated with chunk openings.

In Section 3.4.4 we discussed the possibility that the operationalization of chunk duration may be confounded with the inter-rater agreement. To examine this possibility, we calculated Fleiss' kappa for each extract and correlated the values with random effect slopes by the extract. Fig. 7 shows that the effect of prosody has the strongest relationship with the inter-rater agreement ($r = 0.76$), which is in line with its role as the strongest predictor. In other words, the more listeners can rely on prosody in chunking, the more they agree on where chunk boundaries lie. The correlation of the effect of chunk duration with inter-rater agreement rates is only moderate ($r = 0.4$), indicating that the operationalization of the variable captures additional information. Overall, the convergence between the random effect slopes returned by single-predictor models and inter-rater agreement rates is remarkable.

5. Discussion

To process speech in real time, listeners need to break it down into smaller units. How do they do it? In this study, we examined the effect of five variables: pause duration, prosodic boundary strength, syntactic boundary strength, chunk duration, and word surprisal on chunk boundary perception in linguistically naïve listeners. Mixed effects logistic regression models based on each variable separately showed that all variables were statistically significant in predicting a perceived chunk boundary. This finding supports non-modular approaches to language (Bornkessel-Schlesewsky

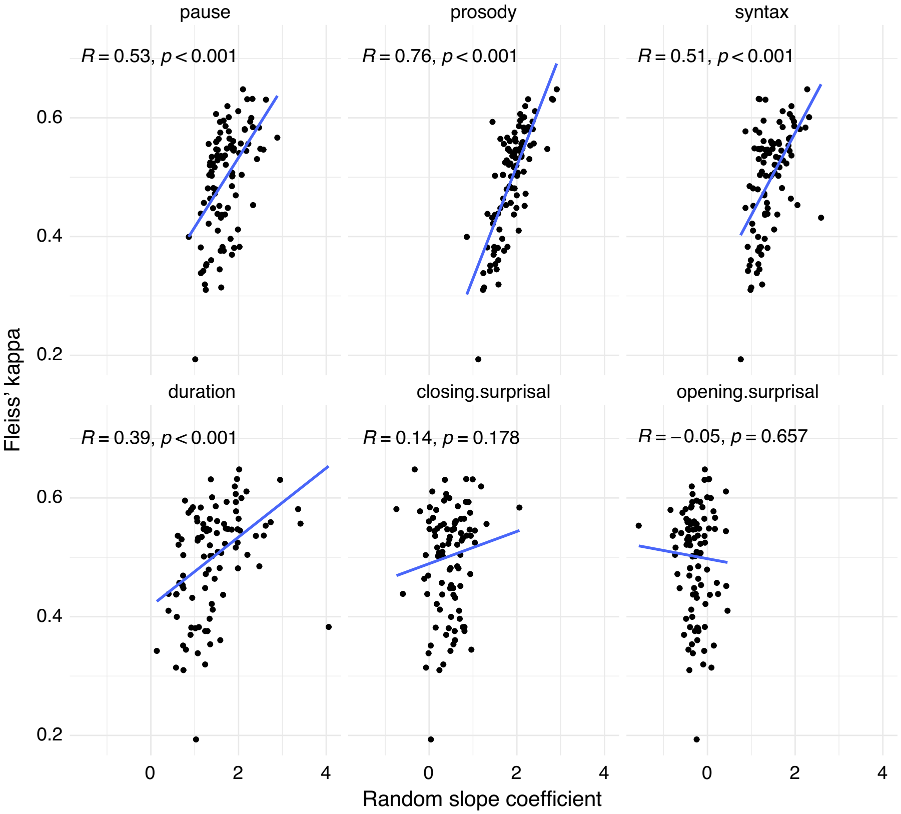


Fig. 7. Scatter plots showing the relationship between inter-rater agreement and the effects of different predictors; each data point represents one extract.

et al., 2016; Goldberg, 2003; MacWhinney, 2012) and suggests that human comprehenders use a variety of cues across different levels of language organization in an integrated manner.

The presence of multiple structurally different cues which perform the same function suggests degeneracy, which is typical of biological systems (Edelman & Gally, 2001). In general, degeneracy provides robustness against variation and perturbations: for example, different brain areas can compensate for each other making cognitive functions resilient to focal brain damage (Noppeney et al., 2004). Similarly, the multiplicity of diverse temporally distributed acoustic cues makes speech robust against noise: for example, the contrast between voiced and voiceless stops can be conveyed by voice onset time, pitch in the following word, the duration of the consonant closure, and loudness (for a review, see Winter, 2014). Cue degeneracy also contributes to the evolvability (Winter, 2014) and learnability (Tal & Arnon, 2022) of language. The finding that pausing, prosody, syntax, and lexical features can all serve to signal chunk boundaries despite being structurally different and simultaneously performing other functions adds to the growing body of literature on functional degeneracy and syntagmatic redundancy of cues in natural language (Leufkens, 2020; Monaghan, 2017; Pijpops & Zehentner, 2022).

https://doi.org/10.1017/langcog.2023.8 Published online by Cambridge University Press

We also found substantial variation in the magnitude of the effects both across listeners and extracts, suggesting that listeners vary in the extent to which they rely on different cues and extracts vary in the extent to which different cues are reliable predictors of chunk boundaries. For example, not all pauses are reliable predictors since for example, a speaker may pause mid-chunk due to lexical search. Similarly, an extract may contain clauses that are too long to serve as chunks. In addition, only moderate slope/slope correlations both across listeners and extracts suggest that listeners may have their individual preferences in the cues they track rather than uniformly track all cues, and extracts may also have specific cues which work best as predictors of chunk boundaries. For example, some listeners may prefer to track prosody, others syntax, and similarly, some extracts may be easier to chunk based on prosody, others based on syntax. Taking a speaker's perspective, the poor syntax of the utterance can be compensated by using prosody, and additional clarity, for example when talking to children, can be gained by marking chunk boundaries with longer pauses.

Listener and extract/speaker variation in the importance of different predictors and the magnitude of their effects is another factor that makes cue degeneracy useful. In fact, chunk boundary cues are not only degenerate but also syntagmatically redundant as evidenced by high intercorrelations between pause duration, and prosodic and syntactic boundary strength: as a result, listeners can rely on any of these cues and still converge on the same chunk boundaries. Earlier, individual variation in the selection and magnitude of the cues listeners attend to was found in prosody perception (Baumann & Winter, 2018; Roy et al., 2017).

As mentioned in Section 3.5, five listeners who did not seem to track any of the cues to the same extent as others were removed from slope/slope correlation analysis. Thus, there are listeners with strong correlations between different effects. Yet, further research is needed to uncover the reasons for this chunking behavior.

A large effect of chunk duration supports the hypothesis that perceptual chunking is affected by a temporal constraint. On average, the duration of a chunk was 2.55 s with SD = 1.2 s. This average falls within 2–3 s which is the bandwidth of delta oscillations and the optimal time-widow for the integration of linguistic information given the time-based working memory constraint (Henke & Meyer, 2021; Roll et al., 2012; Schremm et al., 2015). It thus seems plausible that the temporal constraint regardless of whether the underlying mechanism is working memory capacity or the delta-band oscillations serves to set the temporal window for processing and can help to predict chunk boundaries.

The operationalization of chunk duration in this study raised concerns that it may be confounded with agreement on chunk boundaries. The correlation between the effect of chunk duration on chunk boundary perception across extracts and Fleiss' kappa values of $r = 0.4$ suggests that the variables are sufficiently separate. A strong correlation of inter-rater agreement rates with the effect of prosody on chunk boundary perception suggests that extracts where prosody is a reliable predictor are more 'chunkable.'

The results for the effect of surprisal support the proposed distinction between perceptual and usage-based chunking. If perceptual chunks were learned multi-word units, the words before the boundary should be less surprising because they belong to the ongoing unit, and the words after the boundary should be more surprising since they start a new unit. The models show the opposite results: chunk boundaries associate with higher closing surprisal and lower opening surprisal. However, both effects were very small as well as different in direction across extracts. Thus, the relationship between statistical information and perceptual chunk boundaries requires further research. For example, it is possible that one of the functions of perceptual chunks is to direct attention to items with high informational value. In this study, we used bigram surprisal as a simple measure that could test the hypothesis that perceptual chunks were learned multi-word units. More complex measures should be included in modeling in the future, such as measures based on larger contexts (e.g. 5-g surprisal).

6. Conclusions

The study shows that in chunking up speech in real time, listeners use all the cues investigated. They vary in the extent to which they track different cues and may also prioritize specific cues. Speech materials in turn may vary in the extent to which different cues are reliable predictors of chunk boundaries and may also have specific cues which work best. Variability in processing seems to be facilitated by degeneracy and syntagmatic redundancy of linguistic cues. Thus, if different cues are studied in isolation by artificially constructing linguistic stimuli in speech perception experiments, as is not uncommon, this can undermine the ecological validity of results.

In addition, we have found support for the importance of the temporal constraint in perceptual chunking of speech as well as evidence of the dissociation between perceptual and usage-based chunking. Together these results suggest that perceptual chunking of speech into temporal groups is a distinct process that can inform linguistic theory. It appears that the process is to a large extent determined by cognitive constraints, possibly neural oscillations in the delta band. Language structure may have evolved in a way that meets these constraints.

Acknowledgments. We wish to thank Editor-in-Chief Dr. Bodo Winter and three anonymous reviewers for their very helpful comments and suggestions. We also want to thank Michal Josífko for programming the first version of *ChunkitApp*. This research was supported by grants from the Finnish Cultural Foundation (Grant # 00160622) to Prof Anna Mauranen and from the Kone Foundation (Grant # 202006983) to Dr. Svetlana Vetchinnikova.

Data availability statement. All linguistic materials, data, and code used in the study are openly available in the Open Science Framework repository at https://osf.io/7bta5/.

Competing interest. The authors declare no competing interests.

References

Anurova, I., Vetchinnikova, S., Dobrego, A., Williams, N., Mikusova, N., Suni, A., Mauranen, A., & Palva, S. (2022). Event-related responses reflect chunk boundaries in natural speech. *NeuroImage*, 255, 119203. https://doi.org/10.1016/j.neuroimage.2022.119203

Auer, P. (1999). *Language in time: The rhythm and tempo of spoken interaction*. Oxford University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*, 67, 1–48. https://doi.org/10.18637/jss.v067.i01

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer* (6.2.14). http://www.praat.org/

Bögels, S., Schriefers, H., Vonk, W., & Chwilla, D. J. (2011). Prosodic breaks in sentence processing investigated by event-related potentials. *Language and Linguistics Compass*, 5(7). https://doi.org/10.1111/j.1749-818X.2011.00291.x

Bornkessel-Schlesewsky, I., Staub, A., & Schlesewsky, M. (2016). The timecourse of sentence processing in the brain. In G. Hickok & S. L. Small (Eds.), *Neurobiology of language* (pp. 607–620). Elsevier. https://doi.org/10.1016/B978-0-12-407794-2.00049-3

Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics*, 16(2), 141–158. https://doi.org/10.1093/applin/16.2.141

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide*. Cambridge University Press.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39, E62. https://doi.org/10.1017/S0140525X1500031X

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1–2), 1–31. https://doi.org/10.1080/23273798.2014.963130

Cole, J., Mahrt, T., & Roy, J. (2017). Crowd-sourcing prosodic annotation. *Computer Speech & Language*, 45, 300–325. https://doi.org/10.1016/j.csl.2017.02.008

Culbertson, G., Andersen, E., & Christiansen, M. H. (2020). Using utterance recall to assess second language proficiency. *Language Learning*, 70, 104–132. https://doi.org/10.1111/lang.12399

Dik, S. C. (1997). *The theory of functional grammar: Complex and derived constructions*. Walter de Gruyter.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1), 158–164. https://doi.org/10.1038/nn.4186

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, 81, 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, 85, 761–768. https://doi.org/10.1016/j.neuroimage.2013.06.035

Drury, J. E., Baum, S. R., Valeriote, H., & Steinhauer, K. (2016). Punctuation and implicit prosody in silent reading: An ERP study investigating English garden-path sentences. *Frontiers in Psychology*, 7, 1–12. https://doi.org/10.3389/fpsyg.2016.01375

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, 98(24), 13763–13768. https://doi.org/10.1073/pnas.231499798

Edlund, J., & Heldner, M. (2005). Exploring prosody in interaction control. *Phonetica*, 62, 215–226. https://doi.org/10.1159/000090099

Ellis, N. C. (2017). Chunking in language usage, learning and change: *I don't know*. In M. Hundt, S. Mollin & S. E. Pfenninger (Eds.), *The changing English language* (pp. 113–147). Cambridge University Press. https://doi.org/10.1017/9781316091746.006

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science*, 208(4448), 1181–1182. https://doi.org/10.1126/science.7375930

Ferreira, F. (1993). The creation of prosody during sentence processing. *Psychological Review*, 100, 233–253.

Ferreira, F. (2007). Prosody and performance in language production. *Language and Cognitive Processes*, 22, 1151–1177. https://doi.org/10.1080/01690960701461293

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Frazier, L., Clifton, C., & Carlson, K. (2004). Don't break, or do: Prosodic boundary preferences. *Lingua*, 114 (1), 3–27. https://doi.org/10.1016/S0024-3841(03)00044-5

Gee, J. P., & Grosjean, F. (1984). Empirical evidence for narrative structure. *Cognitive Science*, 8(1), 59–85. https://doi.org/10.1016/S0364-0213(84)80025-7

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66(1–2), 113–126. https://doi.org/10.1159/000208934

Gilbert, A. C., Boucher, V. J., & Jemel, B. (2015). The perceptual chunking of speech: A demonstration using ERPs. *Brain Research*, 1603, 101–113. https://doi.org/10.1016/j.brainres.2015.01.032

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, 15(4), 511–517. https://doi.org/10.1038/nn.3063

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5), 219–224. https://doi.org/10.1016/S1364-6613(03)00080-9

Halliday, M. A. K. (2009). *Language and education*. Continuum.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed). Arnold.

Henke, L., & Meyer, L. (2021). Endogenous oscillations time-constrain linguistic segmentation: Cycling the garden path. *Cerebral Cortex*, 31(9), 4289–4299. https://doi.org/10.1093/cercor/bhab086

Hitch, G. J., Burgess, N., Towse, J. N., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *The Quarterly Journal of Experimental Psychology Section A*, 49(1), 116–139. https://doi.org/10.1080/713755609

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge University Press. https://doi.org/10.1017/9781316423530

Hwang, H., & Steinhauer, K. (2011). Phrase length matters: The interplay between implicit prosody and syntax in Korean "garden path" sentences. *Journal of Cognitive Neuroscience*, 23(11), 3555–3575. https://doi.org/10.1162/jocn_a_00001

Inbar, M., Grossman, E., & Landau, A. N. (2020). Sequences of intonation units form a ~ 1 Hz rhythm. *Scientific Reports*, 10(1), 15846. https://doi.org/10.1038/s41598-020-72739-4

Itzhak, I., Pauker, E., Drury, J. E., Baum, S. R., & Steinhauer, K. (2010). Event-related potentials show online influence of lexical biases on prosodic processing. *NeuroReport*, 21(1). https://doi.org/10.1097/WNR.0b013e328330251d

Kaufeld, G., Bosker, H. R., & Martin, A. E. (2020). Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *The Journal of Neuroscience*, 40(49), 9467–9475.

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLOS Biology*, 16(3), e2004473. https://doi.org/10.1371/journal.pbio.2004473

Kerkhofs, R., Vonk, W., Schriefers, H., & Chwilla, D. J. (2007). Discourse, syntax, and prosody: The brain reveals an immediate interaction. *Journal of Cognitive Neuroscience*, 19(9), 1421–1434. https://doi.org/10.1162/jocn.2007.19.9.1421

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning*, 50(4), 675–724. https://doi.org/10.1111/0023-8333.00143

Leufkens, S. (2020). A functionalist typology of redundancy. *Revista Da ABRALIN*, 19(3), 79–103. https://doi.org/10.25189/rabralin.v19i3.1722

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lüdecke, D. (2022). *sjPlot: Data Visualization for Statistics in Social Science.* R package version 2.8.12. https://CRAN.R-project.org/package=sjPlot

MacWhinney, B. (2012). A tale of two paradigms. In M. Kail & M. Hickmann (Eds.), *Language acquisition and language disorders* (Vol. 52, pp. 17–32). John Benjamins. https://doi.org/10.1075/lald.52.03mac

Männel, C., Schipke, C. S., & Friederici, A. D. (2013). The role of pause as a prosodic boundary marker: Language ERP studies in German 3- and 6-year-olds. *Developmental Cognitive Neuroscience*, 5, 86–94. https://doi.org/10.1016/j.dcn.2013.01.003

McCauley, S. M., & Christiansen, M. H. (2014). A computational model. *Mental Lexicon*, 9(3), 419–436. https://doi.org/10.1075/ml.9.3.03mcc

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, 126(1), 1–51. https://doi.org/10.1037/rev0000126

Meyer, L., Henry, M. J., Gaston, P., Schmuck, N., & Friederici, A. D. (2017). Linguistic bias modulates interpretation of speech via neural delta-band oscillations. *Cerebral Cortex*, 27(9), 4293–4302. https://doi.org/10.1093/cercor/bhw228

Meyer, L., Sun, Y., & Martin, A. E. (2020). Entraining" to speech, generating language? *Language, Cognition and Neuroscience*, 35(9), 1138–1148. https://doi.org/10.1080/23273798.2020.1827155

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97. https://doi.org/10.1037/h0043158

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science*, 9(1), 21–34. https://doi.org/10.1111/tops.12239

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, 14(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Noppeney, U., Friston, K. J., & Price, C. J. (2004). Degenerate neuronal systems sustaining cognitive functions. *Journal of Anatomy*, 205(6), 433–442. https://doi.org/10.1111/j.0021-8782.2004.00343.x

Ostendorf, M., Price, P., & Shattuck-Hufnagel, S. (2005). The Boston University Radio News Corpus. Technical report. Boston University Technical Report No. ECS-95-001, March 1995.

Phipson, B., & Smyth, G. K. (2010). Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 1–12. https://doi.org/10.2202/1544-6115.1585

Pierrehumbert, J. (1980). *The phonetics and phonology of English intonation.* Doctoral dissertation, Massachusetts Institute of Technology.

Pijpops, D., & Zehentner, E. (2022). How redundant is language really? Agent-recipient disambiguation across time and space. *Glossa: A Journal of General Linguistics*, 7(1), 1–41. https://doi.org/10.16995/glossa.8763

Puoliväli, T., Palva, S., & Palva, J. M. (2020). Influence of multiple hypothesis testing on reproducibility in neuroimaging research: A simulation study and Python-based software. *Journal of Neuroscience Methods*, 337, 108654. https://doi.org/10.1016/j.jneumeth.2020.108654

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. (arXiv:2004.09813). arXiv. https://doi.org/10.48550/arXiv.2004.09813

Rimmele, J. M., Poeppel, D., & Ghitza, O. (2021). Acoustically driven cortical δ oscillations underpin prosodic chunking. *ENeuro*, 8(4), 1–15. https://doi.org/10.1523/ENEURO.0562-20.2021

Roll, M., Lindgren, M., Alter, K., & Horne, M. (2012). Time-driven effects on parsing during reading. *Brain and Language*, 121(3), 267–272. https://doi.org/10.1016/j.bandl.2012.03.002

Roy, J., Cole, J., & Mahrt, T. (2017). Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 8(1), 22. https://doi.org/10.5334/labphon.108

Ryan, J. (1969). Grouping and short-term memory: Different means and patterns of grouping. *Quarterly Journal of Experimental Psychology*, 21(2), 137–147. https://doi.org/10.1080/14640746908400206

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274 (5294), 1926–1928.

Schafer, A., Speer, S., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, 29, 169–182. https://doi.org/10.1023/A:1005192911512

Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. In J. J. Ohala (Ed.), *Proceedings of the 14th international congress of phonetic sciences* (pp. 607–610). San Francisco. https://doi.org/10.5282/ubm/epub.13682

Schremm, A., Horne, M., & Roll, M. (2015). Brain responses to syntax constrained by time-driven implicit prosodic phrases. *Journal of Neurolinguistics*, 35, 68–84. https://doi.org/10.1016/j.jneuroling.2015.03.002

Selkirk, E. (1984). *Prosody and syntax: The relation between sound and structure*. MIT Press.

Selkirk, E. O. (1978). On prosodic structure and its relation to syntactic structure. In *Proceeding of the Sloan workshop on the mental representation of phonology*, University of Massachusetts.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). FMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Sinclair, J., & Mauranen, A. (2006). *Linear unit grammar integrating speech and writing*. John Benjamins.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128 (3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Suni, A. (2017). Wavelet Prosody Toolkit. https://github.com/asuni/wavelet_prosody_toolkit

Suni, A., Šimko, J., Aalto, D., & Vainio, M. (2017). Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45, 123–136. https://doi.org/10.1016/j.csl.2016.11.001

Stehwien, S., & Meyer, L. (2021). *Rhythm comes, rhythm goes: Short-term periodicity of prosodic phrasing* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/c9sgb

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

Swerts, M., & Hirschberg, J. (1998). Prosody and conversation: An introduction. *Language and Speech*, 41, 229–233.

Swerts, M., & Hirschberg, J. (2008). Prosodic predictors of upcoming positive or negative content in spoken messages. *Journal of the Acoustical Society of America*, 128, 1337–1344. https://doi.org/10.1121/1.3466875

Tal, S., & Arnon, I. (2022). Redundancy can benefit learning: Evidence from word order and case marking. *Cognition*, 224, 105055. https://doi.org/10.1016/j.cognition.2022.105055

Terrace, H. S. (2001). Chunking and serially organized behavior in pigeons, monkeys and humans. In R. G. Cook (Ed.), *Avian visual cognition*. Comparative Cognition Press.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th international conference on language resources and evaluation (LREC'2012)*(p. 5).

Truckenbrodt, H. (1999). On the relation between syntactic phrases and phonological phrases. *Linguistic Inquiry*, 30(2), 219–255.

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, 142(4), 1976–1989. https://doi.org/10.1121/1.5006179

Vetchinnikova, S., Konina, A., Williams, N., Mikušová, N., & Mauranen, A. (2022). Perceptual chunking of spontaneous speech: Validating a new method with non-native listeners. *Research Methods in Applied Linguistics*, 1(2), 100012. https://doi.org/10.1016/j.rmal.2022.100012

Vetchinnikova, S., Mauranen, A., & Mikušová, N. (2017). ChunkitApp: Investigating the relevant units of online speech processing. In *INTERSPEECH 2017 – 18th annual conference of the international speech communication Association* (pp. 811–812).

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7–9), 905–945. https://doi.org/10.1080/01690961003589492

Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, 19(6), 713–755. https://doi.org/10.1080/01690960444000070

Wickelgren, W. A. (1964). Size of rehearsal group and short-term memory. *Journal of Experimental Psychology*, 68, 413–419. https://doi.org/10.1037/h0043584

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, 36(10), 960–967. https://doi.org/10.1002/bies.201400028

Corpora used

BASE: Thompson, P., & Nesi, H. (2001). The British Academic Spoken English (BASE) Corpus Project. *Language Teaching Research*, 5(3), 263–264. https://doi.org/10.1177/136216880100500305

ELFA (2008). The Corpus of English as a Lingua Franca in Academic Settings. Director: Anna Mauranen. http://www.helsinki.fi/elfa

MICASE: Simpson, R. A., S. L. Briggs, J. Ovens, & Swales, J.M. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

VOICE. (2013). The Vienna-Oxford International Corpus of English (version 2.0 XML). Director: Barbara Seidlhofer; Researchers: Angelika Breiteneder, Theresa Klimpfinger, Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka.

https://doi.org/10.1017/langcog.2023.8 Published online by Cambridge University Press

A. Appendix

ChunkitApp instructions

Humans process information constantly. When we take in information, we tend to break it up quickly into small bits or chunks. We ask you to work intuitively. When you click 'Start', you will listen to a recording and follow it from the text that appears below. Your task is to mark boundaries between chunks by clicking '∽' symbols. One click makes the boundary appear. If you click the symbol again, the boundary will disappear. If you are unsure, put in a boundary rather than leave one out. If you lose the line in the text, stay with the speaker and do not try to go back.

Cite this article: Vetchinnikova, S., Konina, A., Williams, N., Mikušová, N. & Mauranen, A. (2023). Chunking up speech in real time: linguistic predictors and cognitive constraints *Language and Cognition* 15: 453–479. https://doi.org/10.1017/langcog.2023.8