UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Perceptual chunking of spontaneous speech : Validating a new method with non-native listeners

Vetchinnikova, Svetlana; Konina, Alena; Williams, Nitin; Mikušová, Nina; Mauranen, Anna

2022
Elsevier B.V.
http://hdl.handle.net/10138/345043
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, vol. 1, no. 2. https://doi.org/10.1016/j.rmal.2022.100012

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.

Research Methods in Applied Linguistics 1 (2022) 100012

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

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

Perceptual chunking of spontaneous speech: Validating a new method with non-native listeners

Svetlana Vetchinnikova[a,*], Alena Konina[a], Nitin Williams[b], Nina Mikušová[a], Anna Mauranen[a]

[a] University of Helsinki, Finland
[b] Aalto University, Finland

ARTICLE INFO

Keywords:
Speech segmentation
Speech comprehension
Chunking
L2
ELF
Inter-rater reliability
Internal validity
Linear unit grammar

ABSTRACT

Human perception relies on chunking up an incoming information stream into smaller units to make sense of it. Evidence of chunking has been found across different domains, including visual events, music, and dance movement. It is largely uncontested that language processing must also proceed in smaller chunks of some kind. What these online chunks consist in is much less understood. In this paper, we propose that cognitively relevant chunks can be identified by crowdsourcing listener perceptions of chunk boundaries in real-time speech, even if the listeners are non-native speakers of the language. We present a paradigm in which experiment participants simultaneously listen to short extracts of authentic speech and mark chunk boundaries using a custom-built tablet application. We then test the internal validity of the method by measuring the extent to which fluent L2 listeners agree on chunk boundaries. To do this, we use three datasets collected within the paradigm and a suite of different statistical methods. The external validity of the method is studied in a separate paper and is briefly discussed at the end.

1. Introduction

When we listen to somebody speak, we are bombarded with a continuous sound flow that we need to make sense of. Given the limitations of working memory, we are faced with a dilemma: how to make sense of what we hear before it fades, while also taking in new material that is continually coming in. Therefore, linguistic information must be processed rapidly, or else it will be lost. This is what Christiansen & Chater (2016) refer to as the Now-or-Never bottleneck. It is by and large accepted that for rapid processing segmentation of the input is central to speech perception as it is to sensory perception more generally. Domains that also involve continuous unfolding of events in time, such as action or vision, indicate that their perception takes place by segmenting the events into smaller chunks (Gobet et al., 2001; Kurby & Zacks, 2008; Radvansky & Zacks, 2014; Bläsing, 2015; Sridharan et al., 2007), not unlike the perception of static objects (Biederman, 1987). In all, prior research suggests that temporal chunking is a domain-general capacity that humans automatically put to use when exposed to a continuous stream of sensory input, such as when observing others' actions or reading a narrative (Kurby & Zacks, 2016).

Yet, we do not know what these chunks are and how to identify them. Here is a relatively simple extract of speech from academic settings (Example 1):

* Corresponding author.
E-mail address: svetlana.vetchinnikova@helsinki.fi (S. Vetchinnikova).

https://doi.org/10.1016/j.rmal.2022.100012
Received 28 December 2021; Received in revised form 8 April 2022; Accepted 8 April 2022
2772-7661/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

(1)

first of all let me say that I think the dissertation and also as this discussion has shown that this is a piece of work that is extremely relevant and valid not only to an African but also to a wider context and particularly for the third world countries
(the ELFA corpus)

Clearly, to make sense of this extract, we need to chunk it into smaller units. What are these units? In linguistics, chunking is mostly associated with multi-word units of different types, such as collocations (Sinclair, 1991), formulaic sequences (Wray, 2002), lexical bundles (Biber et al., 1999) and constructions (Goldberg, 2006). As in any text, such units are visible in this extract too: *first of all, let me say, I think, wider context, not only X but also Y, third world countries*. However, while previous experience of these units helps a listener understand the extract, this does not imply that they are the units of the segmentation process. Segmenting continuous language input takes place in real time, along the syntagmatic axis. How we as listeners recognize multi-word units is by having encountered them repeatedly, much like we recognize individual words. In linguistic analyses, multi-word units are extracted from large databases as recurrent paradigmatic units. Thus, it seems important to draw a distinction between *perceptual chunking* as online partitioning of incoming signal into temporal groups (see e.g. Gilbert et al., 2015; Rimmele et al., 2020), and *usage-based chunking* as gradual emergence and entrenchment of multi-word units through repeated use (Bybee, 2010; Ellis, 2017; Goldberg, 2006).

Work on statistical learning convincingly shows that even 8-month-old infants are sensitive to the internal associations between sequential elements of the input and learn them spontaneously (Saffran et al., 1996). Thus, input which comes in chunks, such as words or multi-word units, can be successfully chunked by humans simply based on its distributional properties, such as the difference between transitional probabilities within and across chunks. Several computational models implement the process (see also Isbilen et al. 2020; Kidd et al. 2020 for recent experimental evidence). For example, the chunk-based learner (CBL) model (McCauley & Christiansen, 2014, 2019) identifies chunks based on backward transitional probabilities (the frequency of *the cat* divided by the frequency of *cat*) and simultaneously builds an inventory of identified chunks, a "chunkatory" which recognizes already seen chunks in new input. Trained on the CHILDES database (MacWhinney, 2000), the model can segment each individual child's input into relevant syntactic constituents without having the part-of-speech information. It also correctly reproduces 60% of each child's utterances by first chunking the words produced using the chunkatory and then sequencing the identified chunks based on chunk-to-chunk backward transitional probabilities. In effect, the model shows that a child can go a long way in learning the grammar of a language by simply relying on the distributional properties of the input. Another model, Parser (Perruchet & Vinter, 1998), shows that it is possible to extract words from the artificial languages employed in Saffran et al. (1996) even without calculating the transitional probabilities and instead letting the system driven by a limited attention span chunk input randomly and gradually converge towards words based on strengthened representations of co-occurring syllables following the principle of self-organization, common to dynamical systems (Perruchet, 2005; Perruchet & Vinter, 2002). Despite some disagreement about how statistical learning works, computational models do not require understanding because they rely solely on the distributional properties in the input and can therefore also work on artificial languages. While the identification of multi-word units is undoubtedly key to language learning and use, we consider it to be the domain of usage-based chunking. In contrast, perceptual chunking, our focus in this paper, seems to provide a temporal window for further processing of the input (Henke & Meyer, 2021; Rimmele et al., 2020).

If the temporal groups resulting from perceptual chunking are not multi-word units, what are they? Are there other linguistic cues which can drive perceptual chunking? Syntax seems to be one candidate for such a cue. In principle, the whole extract in (Example 1) can be analyzed as one sentence which is clearly too long to serve as a processing chunk, given the limitations of working memory capacity. It can certainly be further divided into smaller constituents, but this is not straightforward, since several different analyses are possible. Grammars and grammarians of different theoretical orientations do not agree on where or how exactly to draw a boundary between units at different hierarchical levels. To take a simple example, *first of all* can be analysed as a linking adverbial (e.g. Biber et al., 1999) or a connective adjunct (Huddleston & Pullum 2002) integrated within the clause, or as extra-clausal (e.g. Dik, 1997) or a "thetical" (Kaltenböck et al., 2011) and placed outside of the clause. The different analyses cannot thus converge for resolving the status of *first of all* – is it a chunk in itself or part of a larger chunk?

Finally, for perceptual chunking listeners may also rely on prosody, but many cognitive studies suggest that prosody is neither necessary nor sufficient for speech segmentation (De Ruiter et al., 2006; Meyer et al., 2017, Itzhak et al., 2010; Ding et al., 2016; Kaufeld et al., 2020). In addition, recent cognitive linguistic theories argue against a modular approach to language and assume that in processing, linguistic information is rapidly integrated from different sources simultaneously (MacWhinney, 2012; Goldberg, 2013; Bornkessel-Schlesewskyet al., 2016). Since prosodic and syntactic units are not fully aligned (Shattuck-Hufnagel & Turk, 1996; Watson & Gibson, 2004; Frazier et al., 2004; Wagner & Watson, 2010), simultaneous reliance on both syntax and prosody should lead to units of processing which differ from strictly syntactic or strictly prosodic units. It would seem, then, that for identifying perceptual chunks we need a method which is independent of theory-driven structural analysis.

A method for exploring listeners' perceptual chunking should also be robust in terms their language status. First and additional language speakers typically show identifiable differences in their language use, but the origin of these differences is debated. Do first (L1) and second language (L2) speakers actually process a given language differently, or have they just had different amounts of exposure to it? One example comes from studies of multi-word chunks. It has often been claimed, since Pawley & Syder, 1983 pioneering study, that learning the phraseology of a language presents a challenge for L2 speakers even at advanced levels of proficiency. The numerous non-standard forms L2 speakers use in multi-word units, such as **on the meantime* (vs. *in)* or **put more attention to (*vs. *pay)* (Yorio, 1989) seem to support this view. Some researchers, notably Wray (2002) and Arnon & Christiansen (2017), propose that this happens because compared to children acquiring their first language, L2 speakers attend to single words, and are therefore not equally sensitive to multi-word units and do not pick them up as easily. In contrast, Mauranen (2012) and Vetchinnikova (2019) suggest that

L2 deviations from L1-like multi-word units arise from holistic but inexact representations in memory, which can be explained by exposure and thereby shallower entrenchment, which leads to more variable use: after all, prepositions and articles, which many L2 learners do not get right, tend to have comparatively little effect on the meaning of a phrase in its context. Moreover, many studies find evidence of usage-based chunking in L2. For example, Gries & Wulff (2005) demonstrate the psycholinguistic reality of constructions for L2 speakers based on evidence of constructional priming. Ellis et al. (2016) test the knowledge and processing of verb- argument constructions in L1 and L2 speakers and find the same effects of construction frequency, type-token distribution, contingency and semantic prototypicality in both groups.

Similar uncertainty concerns perceptual chunking. On the one hand, it is clearly a domain-general process which should be shared by all humans, and on the other hand previous language experience has been found to exert influence on the kind of information that is used for identifying chunk boundaries as well as the weightings assigned to different cues, such as prosody and syntax (see Bates et al., 1982; MacWhinney et al., 1984 for cue validity and the competition model; Bornkessel-Schlesewsky & Schlesewsky, 2019 for the application of cue validity to predictive coding). At the same time, redundancy, ubiquitous at all levels of language, may enable listeners to converge on the same chunk boundaries even if they rely on different cues to identify them.

In this paper we propose that perceptual chunks can be identified by crowdsourcing chunk boundary perceptions from listeners who understand what they hear, irrespective of their native/non-native speaker status. Following Sinclair & Mauranen (2006), we hypothesize that listeners reasonably fluent in a language intuitively chunk speech in approximately the same way and will spontaneously converge on the same chunk boundaries. Their ability to do so rests on understanding the speech they hear. Therefore, the focus on L2 comprehenders allows the inclusion of understanding as a variable: variation in extract comprehensibility and listeners' comprehension skills permits us to explore the relationship between comprehension and chunking. It is reasonable to assume, moreover, for a given language, that if we find high rates of agreement among L2 speakers from different L1 backgrounds, agreement among L1 speakers is at least equally high, if not higher.

Against this background, we selected stimuli and experiment participants to ensure a good level of understanding but leave some space for variation. English as a lingua franca (ELF) contexts, where English is used as a shared language of communication between L2 speakers, seemed to suit this purpose well. Today ELF is unquestionably the predominant mode in which English is used around the world (e.g., Crystal, 2012; Jenkins, 2015) and is therefore of interest in its own right, too. We focused on academic speech situations to maintain a stable context type and high-level language use. The stimuli for our experiments were extracted from corpora of English speech recorded in university settings including both L1 and L2 speakers. Importantly, the speech data was authentic, which raises the ecological validity of the study in comparison to many experimental designs that have been based on artificially constructed clauses, sentences, or strings of these. The experiment participants were university students speaking English as an additional language, that is, speakers who are naturally exposed to the kind of speech we use as our stimuli. As is common to ELF contexts, they came from different L1 backgrounds which should minimize the effect of any particular first language.

To collect data on chunk boundary perception, we asked experiment participants to listen to the stimulus extracts and simultaneously mark chunk boundaries in their transcripts using a custom-built tablet application. In this paper, we examine the internal validity of the method by looking at the consensus between participants by means of inter-rater agreement measures in three datasets of 'chunked' data using several statistical approaches. If the hypothesis is confirmed, a chunk can be defined as a string of words between boundaries reliably identified by naïve listeners, that is, listeners who do not have linguistic training.

In what follows, we discuss the literature on boundary perception in Section 2. In Section 3, we turn to the concept of consensus and consider the suitability of different statistical measures commonly used to estimate inter-rater reliability. In Section 4.1, we present the task we used to collect listener perceptions and discuss the caveats involved. Sections 4.2–4.4 provide details about the speech materials, experiment participants and experiment setting. In Sections 4.5 and 4.6, we explain how we used Fleiss' kappa and permutation-based tests to provide different perspectives on listener consensus. In Section 5 we present the results, discuss them in Section 6 and draw conclusions in Section 7.

2. Boundary perception

Discussing segmentation tasks in event perception research, Kurby & Zacks (2008: 72) write: "How can a researcher discover when a person perceives that a new event has begun? One simple but surprisingly powerful answer is simply to ask them, usually by having them press a button". Yet, despite the simplicity and intuitive appeal of this approach, to our knowledge there is not much research that employs it, especially in linguistics. In event perception research, the segmentation task was first introduced by Newston (1973). Typically, research participants are asked to watch a short movie and press a button "whenever, in [their] judgement, one meaningful unit of activity ends and another begins", as Radvansky & Zacks (2014: 81) put it. The authors note that participants often find the instructions confusing but almost always perform well on the task and emphasise that they "produce strikingly regular data". For example, Speer et al. (2003) had their participants segment the same movies twice with a one-year gap in-between and found high test-retest correlations as well as significantly greater than chance intra- and interindividual agreement. Kurby & Zacks (2008) point out that good interindividual agreement and reliability provide evidence of the validity of the task and the cognitive reality of the construct the task is trying to tap into, that is, the automaticity of (event) segmentation. Clearly, this should apply to language data too.

Sinclair & Mauranen (2006) used boundary perception in developing a grammatical model of a radically new type. They argued that a linear, bottom-up grammar built on intuitively perceived increments can better account for real-time cognitive processing constraints and bridge the gap between speakers' experience of language and linguists' descriptions of it. Thus, to avoid imposing an analyst's predetermined categories on natural speech data, they separated increment identification from the linguistic analysis.

They first used their own intuitions, independently from each other, to chunk up stretches of language, and only then developed an analytical framework for those chunks. In this way, the chunking stage was intuitive and involved attention to perceived boundaries between chunks, without concern for what they consisted of. No preconditions were set on the shape or size of the units. Sinclair & Mauranen (2006) moreover suggested that anyone fluent in the language would chunk it up in approximately the same way. However, in the original model the chunking stage relied on only two coders, the authors themselves, and did not involve larger data collection on listener boundary perception. It would therefore be important to put this to test, and to find out how far this assumed consensus holds for fluent speakers of English, in other words, how much agreement we can detect in their perceptual chunking.

In this paper, we propose collecting listener perceptions of chunk boundaries from a large number of naïve listeners, or, in effect, crowdsourcing chunk boundary identification. We suggest that the internal validity of the method can be assessed by measuring interindividual agreement on chunk boundaries. The quantification of the extent of agreement over multiple raters is a statistically challenging question, not least because of the peculiar distribution of the data. In the next section, we consider available measures of inter-rater reliability.

3. Inter-rater reliability

The obvious way to measure agreement in the chunking task is to adopt one of the measures of inter-rater reliability widely used in psychology and applied linguistics. Inter-rater reliability measures broadly divide into estimates of agreement and estimates of consistency (see e.g. Stemler, 2004). While the former assume that raters should come to an exact agreement about the interpretation and application of a given rating scheme, the latter only require them to behave consistently with regard to their own interpretation of the construct but not necessarily agree on the specific ratings with each other; in other words consistency estimates measure how the ratings "hold together to measure a common dimension" (Stemler, 2004: 4). While in principle both types of measures are applicable to our data, neither of them is entirely unproblematic. In this paper, measures of agreement are more relevant to our objectives, and we focus on those.

The simplest agreement estimate is observed agreement or percent agreement, that is, the number of rater-rater pairs in agreement as a proportion of the total possible number of rater-rater pairs. However, percent agreement might be misleading because some agreement would arise by chance alone. Cohen's kappa (Cohen, 1960) improves on this measure by explicitly correcting for agreement due to chance. It is calculated as the difference between the observed agreement and chance agreement divided by the agreement attainable above chance and ranges from -1 to 1. 1 indicates perfect agreement, 0 random agreement and -1 complete disagreement (Hallgren, 2012). While Cohen's kappa calculates agreement for two raters, Fleiss' kappa is suitable for multiple raters, but in essence it is the same statistic.[1]

The literature provides recommended benchmarks for interpreting kappa statistics. Landis & Koch (1977) suggest that <0 should be considered 'poor agreement', 0.01-0.2 'slight agreement', 0.21-0.4 'fair agreement', 0.41-0.6 'moderate agreement', 0.61-0.8 'substantial agreement' and 0.81-1 'almost perfect agreement'. Plonsky & Derrick (2016) conducted a meta-analysis of 2,244 reliability estimates from 537 articles published in the field of L2 research, which included 369 percent agreement estimates and 145 Cohen's kappa estimates and propose more conservative benchmarks based on the distribution of values reported in the field. They suggest that the 25th percentile can serve as the threshold for a minimally acceptable estimate and 75th percentile for a high-reliability estimate. However, the median values they found for percent agreement and Cohen's kappa are surprisingly high with a very narrow interquartile range (IQR) too: 0.93 (median) and 0.08 (IQR) for percent agreement and 0.87 (median) and 0.17 (IQR) for Cohen's kappa. They acknowledge that the values might be inflated since researchers are more likely to report higher reliability estimates. This explanation appears to be plausible since in L2 research if interrater reliability is low, raters can and should be re-trained to achieve higher reliability. In all, while such benchmarks provide a useful perspective, there are two caveats that need to be taken into account when applying them to our data.

First, the concept of inter-rater agreement hinges on the ability of raters to interpret and apply the guidelines they are supplied with in a similar way, such as a scoring rubric in a language test (see Grabowski & Oh, 2018) or a coding scheme in linguistic annotation (e.g. Larsson et al., 2020). It is also common to train the raters and make sure they achieve consensus, which can also contribute to the very high estimates of reliability obtained by Plonsky & Derrick (2016). In contrast, we provide only very minimal guidelines (see Section 4.1). Thus, if we still find agreement between our annotators, the agreement is not due to the quality of the guidelines or training but rather due to the cognitive reality of the construct we are trying to tap into. For example, in an event segmentation study Speer et al. (2003) found the mean intra-rater percent agreement of only 0.38 (SD = 0.16) and a correspondingly low inter-rater percent agreement of 0.28 (SD = 0.6) which they nevertheless interpret as significant and noteworthy.[2]

Second, as will be clear from Section 5.1, in our data we are faced with a vastly uneven distribution of boundary markings and non-markings at the individual level and a grossly non-normal distribution of aggregated boundary markings at the group level. This makes the application of traditional measures of inter-rater reliability challenging. In fact, it is long known that both Cohen's kappa and Fleiss' kappa are strongly affected by prevalence, i.e., imbalance in the number of responses across categories, and bias, the difference between raters in how often they assign a case to a specific category (Feinstein & Cicchetti, 1990; Byrt et al., 1993; Hallgren, 2012). Prevalence typically produces underestimation of consensus, and bias overestimation. In our data prevalence is

[1] There is a slight difference in how the two calculate chance agreement, since Fleiss' kappa is a generalisation of Scott's pi rather than Cohen's kappa directly. But in any case they share very similar properties.

[2] Intra-rater agreement was calculated as the proportion of segment boundaries identified in one session that were also identified as boundaries in the second session. Likewise, in calculating inter-rater agreement, it seems only agreement on boundaries was taken into account.

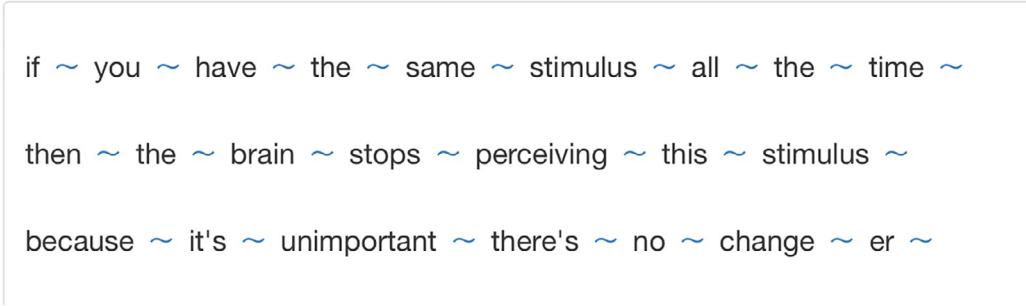


Fig. 1. User interface, *ChunkitApp*.

clearly present as a boundary remains unmarked ten times more often that it is marked, but for the same reason bias can only be very small.[3]

There have been attempts to solve the problem. For example, Gwet (2008, 2014) proposed a new measure, Gwet's AC1, which introduces a different definition of chance agreement but retains all other parts of the formula. However, the values of Gwet's AC1 for our data were very close to the values of percent agreement suggesting that, at least for our data, chance agreement is not sufficiently accounted for. In fact, non-marking in the chunking task is more likely in cases of uncertainty, so the larger the proportion of non-markings in an extract, the more the estimate of agreement should be corrected for chance. In a similar vein, Byrt et al. (1993: 428) point out that it is not necessarily bad that Fleiss' kappa is dependent on prevalence since unequal distribution of data across categories gives more room for chance agreement. At the same time, they emphasise that the effect of prevalence can be very large and the obtained kappa coefficients cannot be directly compared between situations where the prevalence is different. In fact, they suggest that prevalence, bias and observed agreement are three different components of kappa statistics all of which should be reported and taken into account in the interpretation of the obtained value. This is the recommendation we will follow in this paper. Given the difficulties involved in interpreting Fleiss's kappa coefficient, we will also compare the original Fleiss' kappa value against a set of null Fleiss' kappa values representing 'no consensus'. These methods are described in Sections 4.5 and 4.6.

Following the above discussion of the literature on boundary perception and inter-rater reliability measures, in probing the internal validity of crowdsourcing chunk boundary perception data from naïve listeners, our specific research questions are:

(1) To what extent do fluent L2 listeners who understand what they hear agree on chunk boundaries?
(2) How much does listener agreement vary across speech extracts?
(3) To what extent does extract quality influence listener agreement?

Importantly, we focus solely on the perceptual end of chunking without making any assumptions about speaker chunks. Speaker produced and listener perceived chunks may or may not correspond but this is beyond the scope of the present study.

4. Data and methods

In this paper, we use three datasets (A–C). In all three datasets, participants recruited from a university student population (Section 4.3) were asked to perform the same chunking task (Section 4.1). The speech materials they listened to were all extracted from academic spoken corpora but following different selection principles (Section 4.2). The transcripts of the extracts, sample audio files, and responses collected from the experiment participants are publicly available at https://osf.io/7w4k9.

4.1. ChunkitApp and the chunking task

To collect chunking data, we designed a custom web-based tablet application *ChunkitApp* (Vetchinnikova et al. 2017, see https://osf.io/7w4k9 for open source *ChunkitApp 2.0*). The application displays transcripts of audio clips at the same time as they play in participants' headphones and allows marking chunk boundaries between words in real time. All orthographic words in the transcript are separated with an interactive tilde symbol (~): by tapping it, participants can insert a boundary or remove it if they change their mind (see Fig. 1).

The instructions participants receive are minimal. They are asked to listen to each recording while following the transcript and "mark boundaries between chunks by clicking '~' symbols" (see Appendix for the full text of the instructions).

[3] The bias index (BI) for Cohen's kappa suggested by Byrt et al. (1993) can be calculated as the number of times participant A marks a boundary, but participant B does not minus the number of times participant A does not mark a boundary, but participant B does divided by the total number of ratings. Byrt et al. does not generalise the formula for multiple raters. However, it is clear that our data has a very small BI since on average the denominator is going to be more than ten times bigger than the numerator. The prevalence index (PI) is the difference between the total number of boundary markings minus the total number of non-markings divided by the total number of ratings.

Table 1
Description of the datasets.

Set	n of extracts	n of words	n of potential boundaries	n of participants	source corpora
Set A	98	5,293	5,195	53	ELFA, VOICE, MICASE
Set B	97	5,237	5,140	51	
Set C	66	4,865	4,799	45	ELFA

The notion of a chunk is not explained. Each recording is played once only. While it is playing, participants can click on the 'pause' button, but the transcript does not show when the recording is paused. These features are designed to encourage participants to make fast online decisions relying solely on their intuition and restrict opportunities for conscious analysis or backtracking. In the experiments reported in this paper, each recording was followed by a self-evaluation or a true-false comprehension question. The questions were used to measure comprehension and to keep the participants focused on the task.

While mimicking natural speech processing would ideally require an aural only presentation, no better technical solution has been available so far. Without the transcript, we would not know where exactly listeners intended to place a boundary even if we time their responses, as they can click slightly earlier, in anticipation of the coming boundary, or slightly later, as a post hoc realisation of the past boundary. Yet, while the availability of a matching transcript may enhance speech processing, the participants clearly attend to the audio as they show sensitivity to acoustic properties of chunk boundaries, such as pause length and prosody (Anurova et al., 2022). Also, Dąbrowska's (2020) review convincingly argues that orthography affects online speech processing in anyone who has acquired literacy, that is, even without simultaneous visual presentation. Finally, in a follow-up listening-only brain imaging experiment, we find evidence that chunk boundaries and non-boundaries identified through ChunkitApp elicit a markedly different response in the brain (Anurova et al., 2022).

The output of the chunking task from each participant is a sequence of orthographic words with marked (represented by 1) and unmarked (represented by 0) boundaries between them (Example 2).

(2)

or (1) I (0) mean (1) there's (0) a (0) school (0) down (1) not (0) very (0) far (0) from (0) here (1)

Since every space between two orthographic words can in principle be marked, the number of *potential boundaries* in each transcript is the number of words minus 1 (there is no space after the last word an extract). The group value for each potential boundary is simply the number of times it was marked in total in a given experiment, which we call its *boundary frequency*. For example, if we have 45 participants and 15 of them marked the boundary (1 × 15) while the remaining 30 did not mark a boundary (0 × 30), the boundary frequency of the given boundary is 15. Boundary frequencies can be examined from the point of view of inter-rater agreement, statistical significance and *boundary strength*.

There are previous studies where a similar research paradigm was used, even though for a different purpose. Jennifer Cole and colleagues (e.g. Cole et al., 2017) explore the usability of a similar task to crowdsource prosodic annotation. They use Language Markup and Experimental Design Software (LMEDS; Mahrt, 2016) for their rapid prosody transcription method (RPT) where they ask untrained annotators to mark words for prosodic boundary and prominence. As in our study, annotation is performed in real time while listening to the corresponding audio recording but in their experiment, annotators listen to each recording twice, first marking boundaries and then prominences. It seems that the main difference between the two applications is that RPT aims to elicit naïve prosodic analysis, while the chunking task strives to tap into what listeners naturally do when they listen to speech, in effect manoeuvring participants away from making conscious decisions which can be based on whatever received ideas about language structure they might have. Working on a tablet is also much faster than using a computer and a mouse, or paper and pencil for that matter, and therefore should give a better window into the online processing we are interested in.

4.2. Speech materials

To enhance the ecological validity of the experiment, the speech materials for all three datasets were sourced from three corpora of authentic naturally occurring academic speech recorded in university settings: the Corpus of English as a Lingua Franca in Academic Settings (ELFA, 2008), the Vienna-Oxford International Corpus of English (VOICE, 2013) and the Michigan Corpus of Academic Spoken English (MICASE, Simpson et al., 2002). ELFA and VOICE are corpora of English used as a lingua franca, in which most of the speakers are non-native speakers of English, and MICASE is an L1 English corpus. Together the corpora reflect the kind of English people are typically exposed to in today's highly international universities. Thus, we do not make a categorical distinction between speech produced by native and non-native speakers and instead focused on clarity and comprehensibility of their speech when selecting the extracts for the stimuli, as detailed below.

For each of the three datasets, we selected a number of short extracts based on the transcripts and cut the corresponding audio clips from corpus files (see Table 1). The selection procedures were different, which allowed us to examine how the quality of the extracts influenced agreement on chunk boundaries. For sets A and B, we followed a set of explicit criteria aiming to collect a homogeneous set of extracts which would be clear, relatively fluent and easy to understand outside the context of the speech event they come from. To identify such extracts in the corpora, we first automatically retrieved stretches of speech longer than 50 words which did not contain unintelligible or unfinished words, laughter, long pauses, overlapping speech, speaker changes, or frequent hesitations

Table 2
Participants across the three datasets.

Dataset	Set A	Set B	Set C
N of participants	53	51	45
Female, %	66%	71%	69%
Right-handed, %	94%	86%	91%
N of different L1s	19	16	13
Finnish L1, %	58%	59%	58%
N of extracts with self-evaluation questions	72	72	66
% of extracts understood, M and SD	97%, 5%	96%, 11%	95%, 7%
N of extracts with true/false questions	25	25	0
Comprehension question score, M and SD	89%, 7%	89%, 9%	N/A

or repetitions.[4] We also controlled for specialized and low-frequency vocabulary. At the next stage, five researchers worked with the automatically generated list and selected extracts which they agreed were sufficiently clear and fluent. Agreement was reached through negotiation. An extract was considered meaningful outside of context if it was possible to formulate a general true/false comprehension question about its main idea. Finally, to achieve even audio quality across all extracts, we recruited a speaker who reproduced them as close to the original audio clips as possible. Having all extracts reproduced by the same speaker also controls for the possible effect of sociolinguistic information, such as age, gender and accent on chunk boundary perception. The new audio files were recorded in an acoustically shielded studio at the phonetics laboratory of the University of Helsinki. For set C, we used a random sample of extracts representative of the corpus and the original audio clips, which thus contained the speech of different speakers recorded in different settings.

4.3. Participants

All participants come from the student population of the University of Helsinki. In total we recruited 149 students aged 20-39 from all fields of science except linguistics to avoid possible bias (Table 2). Females were slightly overrepresented, but gender is not a critical variable in this study. Most participants were right-handed. None reported diagnosed dyslexia. Importantly, all students were L2 speakers of English coming from a variety of different L1 backgrounds which is typical of ELF contexts. Finnish was the most common L1.

As mentioned in the introduction, we expected the participants to understand the speech extracts used in the experiment because this is the kind of English they are normally exposed to as university students. To probe their understanding, we asked them to answer a comprehension question after each extract. A quarter of all extracts in sets A and B were supplied with true/false comprehension questions, piloted with a separate group of participants. To avoid experiment fatigue, the rest of the extracts were followed with a simple self-evaluation question: “Did you understand what the speaker was saying?” with three answers available: yes/no/roughly. In set C, we used self-evaluation questions only. As Table 2 shows, on average the participants reported that they understood or roughly understood 95–97% of extracts and answered correctly 89% of true/false comprehension questions.

4.4. Experiment setting

The participants were invited to the experiment in small groups and provided with iPads which contained all instructions guiding them through different phrases of the experiment. In addition to the chunking task itself, the task battery included a background questionnaire, a short proficiency test (an elicited imitation task or a yes/no vocabulary test), a feedback form and a few other tasks not relevant here. The experiment was overseen by a researcher and took up to two hours including a coffee break. The participants could take additional breaks at any time. All data was handled completely anonymously as the participants were not asked to provide their name during the experiment. Informed consent was obtained, and all participants received a movie ticket as a compensation for their time and effort.

4.5. Measuring consensus with Fleiss’ kappa

As discussed in Section 3, Fleiss’ kappa appears to be the most suitable method of measuring listener consensus in our data. It is applicable to data collected from multiple raters and it accounts for chance agreement. However, it is seriously affected by prevalence, i.e., an uneven distribution of ratings between categories: in our case a much larger number of 0s than 1s (Feinstein & Cicchetti, 1990; Byrt et al., 1993; Hallgren, 2012). For this reason, Fleiss’ kappa values obtained from data characterised by high prevalence cannot be directly compared to the available benchmarks or values obtained in other studies.

Thus, to interpret Fleiss’ kappa, we report the observed agreement and the prevalence index along with the kappa coefficient itself. In addition, we use a permutation-based method, wherein we compare the estimated Fleiss’ kappa value with a null distribution of Fleiss’ kappa values calculated for datasets representing ‘no consensus’, but with exactly the same distribution of non-markings and

[4] Some of these criteria were dictated by the requirements of a parallel brain imaging experiment we report in Anurova et al. (2022).

Table 3
Observed agreement and Fleiss' kappa across the datasets.

Set	Observed agreement	Fleiss' kappa κ	CI	Prevalence Index (PI)
Set A	0.921	0.536	[0.535, 0.536]	0.81
Set B	0.916	0.486	[0.486, 0.487]	0.82
Set C	0.897	0.429	[0.428, 0.430]	0.80

markings as in the original dataset. The null distribution is generated by random resampling (without replacement) the set of responses of each participant, and estimating Fleiss' kappa on the resulting dataset. This is done 1,000 times to produce a distribution of 'no consensus' Fleiss' kappa values.

4.6. Measuring statistical significance of boundary frequencies

Fleiss' kappa provides an overall measure of agreement, for a specific extract or an entire dataset. Testing boundary frequencies for statistical significance is a way to assess agreement after each word: simple percentage of participants who mark a boundary after a given word does not take into account chance agreement or the overall distribution of boundary markings.

In effect, in the chunking task a participant makes a binary choice between a boundary and a non-boundary after each word. Thus, at the group level, some boundary frequencies are possible even if all participants are marking boundaries at random. To compare, if one flips a fair coin twice, there is a 50% chance to get 1 head and 1 tail, 25% chance to get 2 heads, and 25% chance to get two tails. How many heads/tails or boundaries/non-boundaries can we expect to get if we flip a fair coin 45 times, which is the number of participants in set C? If we assume that the participants of the chunking task are acting at random, we should expect a binomial distribution of boundary frequencies. Yet, in our case the coin is not 'fair' since on average a non-boundary is ten times more probable than a boundary and the ratio between boundaries and non-boundaries also varies across participants. Thus, to obtain the null distribution and determine the statistical significance of boundary frequencies, we used a permutation test.

For each dataset, we permuted individual markings (i.e., 0s and 1s for each individual) one million times (with replacement). For each boundary position we compared the observed boundary frequency with one million permuted boundary frequencies and calculated a two-tailed *p*-value. To avoid zero *p*-values in cases where the observed or a more extreme boundary frequency did not occur in the permutations as not all possible permutations were run, we defined *p* as the upper bound $p_u = (b + 1)/(m + 1)$ where *b* was the number of times permuted boundary frequency was equal or more extreme than the observed and *m* was the number of permutations (Phipson & Smyth, 2010; Puoliväli et al., 2020). The number of permutations was high enough to make the *p*-values for the same boundary frequency across different boundary positions approximately the same. However, since *p* was computed for each boundary, we applied the Benjamini-Hochberg false discovery rate (FDR) procedure (Benjamini & Hochberg 1995) to account for multiple comparisons. We used MultiPy package for Python (Puoliväli et al., 2020). The critical level was set at $\alpha = 0.05$. Boundary frequencies which were higher than expected by chance were considered statistically significant chunk boundaries. Boundary frequencies which were lower than expected by chance were considered statistically significant non-boundaries.

In sum, the method determines statistical significance of boundary markings after each word as well as allows us to distinguish between agreement on boundaries (upper tail of the distribution) and agreement on non-boundaries (lower tail). Together, the Fleiss' kappa analysis and the analysis of the statistical significance of boundary frequencies give a more complete picture of consensus than either analysis could give separately.

5. Analysis and results

5.1. Distribution of boundary markings and boundary frequencies

On average an experiment participant marks 9–10% of all potential boundaries (for sets A–C: mean = 10%, 9%, 9%, SD = 5%, 4%, 4%, see Fig. 2). In other words, some participants mark shorter chunks, some longer, but on average the length of a chunk as perceived by an individual participant is 9 to 10 words.

At the group level, the distribution of boundary frequencies seems to closely approximate a Zipfian distribution, like many other phenomena in language (Piantadosi, 2014; Ellis et al., 2016). About half of all boundaries in each of the datasets (44–58%) are not marked by any of the participants (zero boundary frequency) and 14% to 25% are marked by just one participant. Fig. 2 shows log-log plots of the relationship between boundary frequency rank and its frequency in each of the datasets. The labels on the axes are shown in original units for ease of interpretation. Note that boundary frequency rank equals boundary frequency plus one, since its minimal value is zero. The plots are approximately linear up to the rank of about 20 as a locally-smoothed regression line (LOESS) indicates.

5.2. Listener consensus

All three datasets revealed strong observed agreement of boundary markings across subjects with values around 90% agreement (Table 3). In contrast, the values of Fleiss' kappa correspond to 'moderate agreement' according to the generic benchmarks (Landis & Koch, 1977). However, as discussed in Section 3, generic benchmarks cannot be applied to our data due to very high prevalence. At

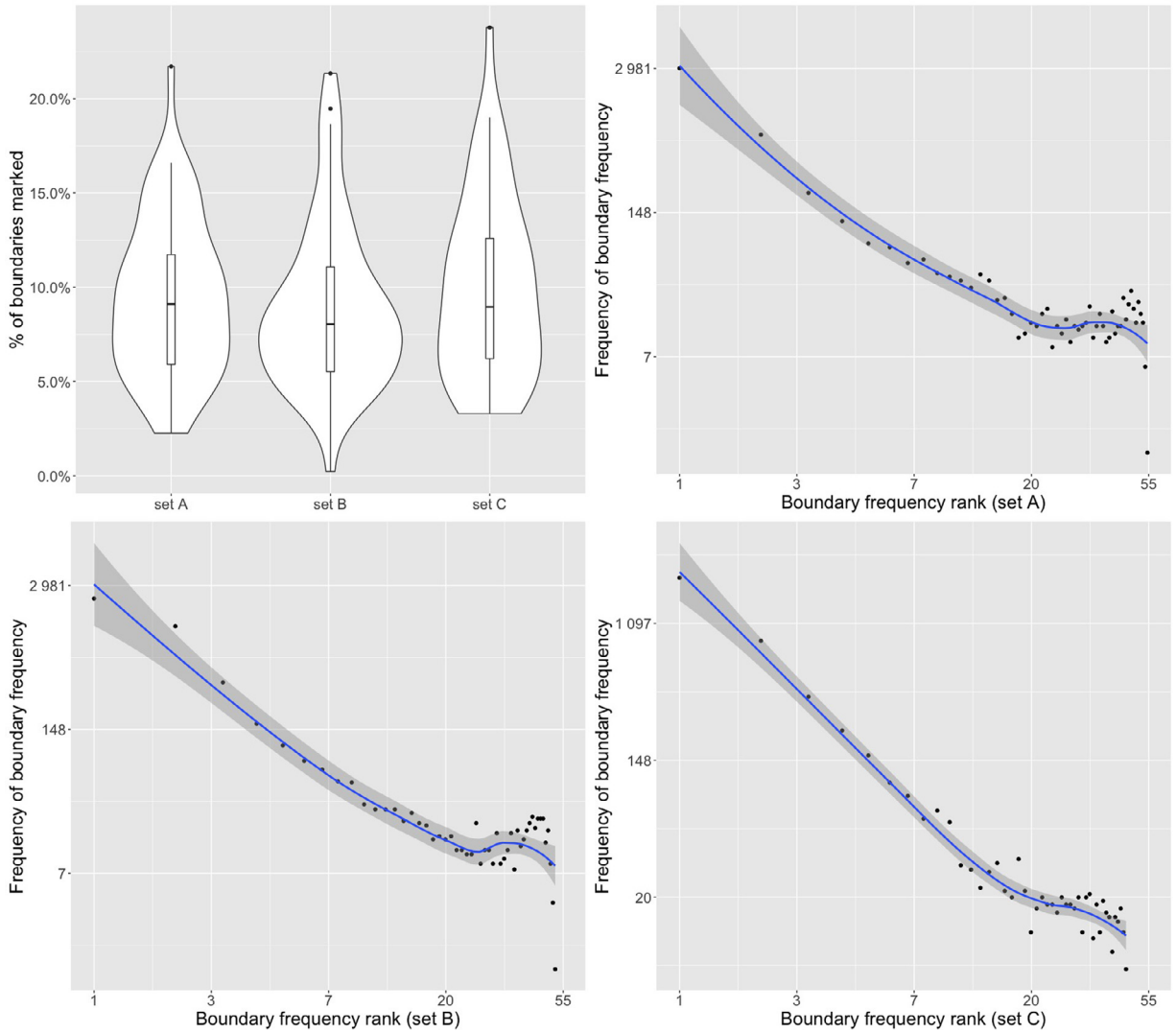


Fig. 2. Proportions of boundaries marked by experiment participants and distribution of boundary frequencies across the sets.

the same time, since the prevalence is constant across the datasets, the obtained values can be meaningfully compared to each other, as well as to other studies where the data has similar properties.

To get another perspective on the obtained kappa values, we generated the null distribution of Fleiss' kappa values representing 'no consensus' for dataset C which has the lowest value (0.429). We found the Fleiss' kappa value for dataset C to be appreciably higher than the mean of the null distribution of Fleiss' kappa values for this dataset ($z = 972.2$, $p < 0.0001$). The null distribution had very low values between -0.002 and 0.0009 with a mean value of -0.0005 (Fig. 3). Thus, we conclude that the original Fleiss' kappa value we obtained is highly unlikely to have arisen due to chance.

While Cole et al. (2017), who explore the usability of crowdsourced prosodic annotation with a similar paradigm (see Section 4.1), do not mention prevalence, it is likely that prosodic boundaries are distributed in a similarly disproportionate way and our findings are therefore comparable. Interestingly, the kappa values they report for agreement on prosodic boundaries in different cohorts of annotators are slightly lower overall. American English speakers working in lab-based conditions reached the agreement of κ=0.51, while American English and Indian English speakers recruited through Amazon Mechanical Turk obtained κ=0.43 and κ=0.23, respectively. One reason for the lower agreement rate could lie in the smaller number of annotators (32), but Cole et al. also show that after a minimum of 20 annotators, the kappa values stabilise and do not increase much anymore. At the same time, many other conditions of Cole et al.'s experiments were more favourable for reaching consensus than in this study: the annotators listened to the recording twice, the recordings were shorter (13–24 s) and there were fewer of them (16 extracts, 931 words in total). All three cohorts of annotators were also L1 speakers of English, and in the case of American English speakers listened to the speech of the variety most familiar to them since the materials were taken from a corpus of American conversational speech. Thus, despite having more room for error and variability in their backgrounds, our participants nevertheless seem to agree more.

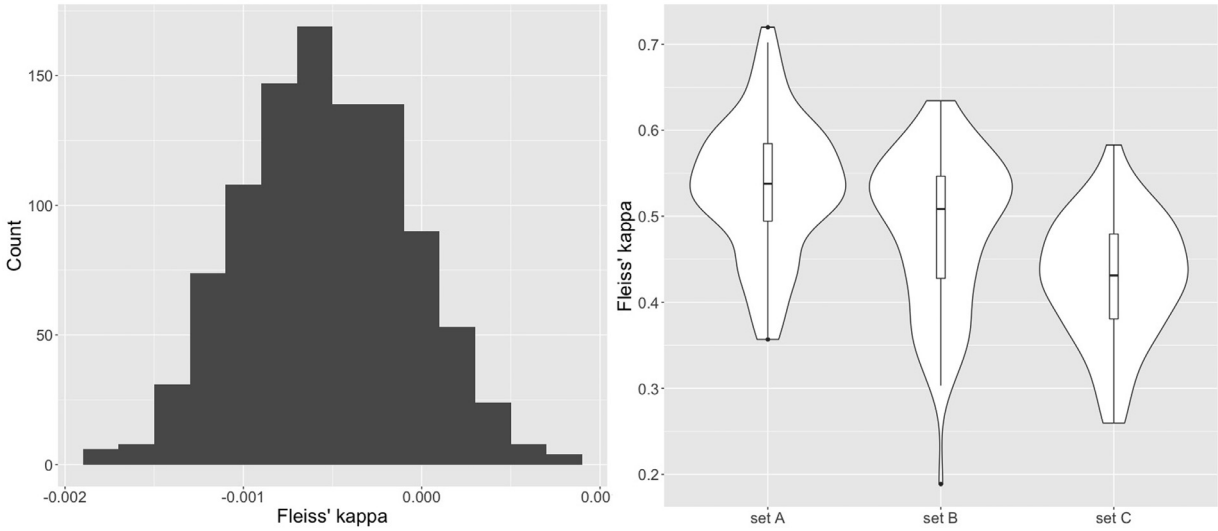


Fig. 3. Null distribution of Fleiss' kappa values and observed Fleiss' kappa across extracts in sets A–C.

The rate of agreement reached across the three datasets in our study varies. Given the differences in the selection of extracts, it is not surprising that the average kappa value in set C is the lowest. The difference in the overall rate of agreement between sets A and B is less expected. The violin plots in Fig. 3 showing the distribution of kappa values across the extracts suggest that the datasets are indeed different. There is also slightly more variance in set B (SD for sets C, A and B is .072, .078 and .088 respectively), though Levene's test for homogeneity of variances conducted on the three datasets does not return a statistically significant result ($p = .076$).

A one-way ANOVA confirms that the difference in kappa values between the datasets is significant, $F\ (2, 258) = 34.52$, $p < .001$, $\eta^2 = .211$. Planned contrasts show that as expected set C is significantly different from sets A and B combined ($t\ (258) = -7.098$, bootstrapped[5] $p < .001$), but that set A is also significantly different from set B ($t\ (258) = 4.301$, bootstrapped $p < .001$). Still, the effect size of the difference between set C and sets A and B is larger than the effect size of the difference between set A and set B: for the first contrast Cohen's $d = -1.008$, 95% CI [-1.299; -.715] [6] and for the second $d = .593$, 95% CI [.306; .878], which can be interpreted as a large and a small to medium effect according to Plonsky & Oswald's (2014) recommendations of benchmarks for between-group contrasts in applied linguistic/L2 research. Thus, it seems that the selection and quality of the extracts have an impact on listener consensus on chunk boundaries: when listening to fluent, easy-to-understand extracts, listeners converge on the same chunk boundaries to a larger extent.

To gain a better understanding of whether Fleiss' kappa actually captures listener consensus in our data, we compared the extracts with the highest and the lowest kappa values in set B. The line charts in Fig. 4 display the sequence of boundary frequencies starting from the beginning of each extract.

It is evident that while extract 191 has clearly defined, almost entirely flat troughs and high peaks representing nearly total agreement on a non-boundary or a boundary, extract 123 is more undulating with no single zero boundary and most boundaries marked by one to ten listeners. A look at the syntax of the extracts suggests that things like a prepositional phrase added as a noun post-modifier (*websites for receiving ideas, complaints…*), listing (*ideas, complaints and so on*), a series of noun pre-modifiers (*citizens' central information centre*) and coordination of verb phrases (*is producing data and distributing data*) in extract 123 might have produced uncertainty about the location of chunk boundaries. This suggests that extracts might vary in how easy or difficult they are to chunk, for example on account of their syntax. "Chunkability" might also interact with comprehensibility, but this requires further research.

5.3. Statistically significant boundaries and non-boundaries

The results of the permutation tests showed that boundaries marked by more than 10 participants in set B, more than 11 in set A, and more than 12 in set C are statistically significant chunk boundaries, that is, boundaries which cannot be attributed to chance behaviour (see Section 4.6). Boundaries not marked by any participant (boundary frequency of 0) are statistically significant non-boundaries in sets A and B but not in C. Fig. 5 shows the probability distribution of boundary frequencies for set A as an example.

The analysis of the statistical significance of boundary frequencies gives rise to a number of observations. First, as discussed in Section 5.1, more than half of boundaries in all three datasets are not marked by any participants. Permutation tests show that the probability of a zero boundary frequency is very low, which supports the interpretation that the data shows notable consensus. The

[5] All bootstraps reported in this paper are based on 10,000 samples.

[6] All Cohen's *d* values reported in this paper use the pooled standard deviation for the groups involved in the contrast and the Hedges' correction.

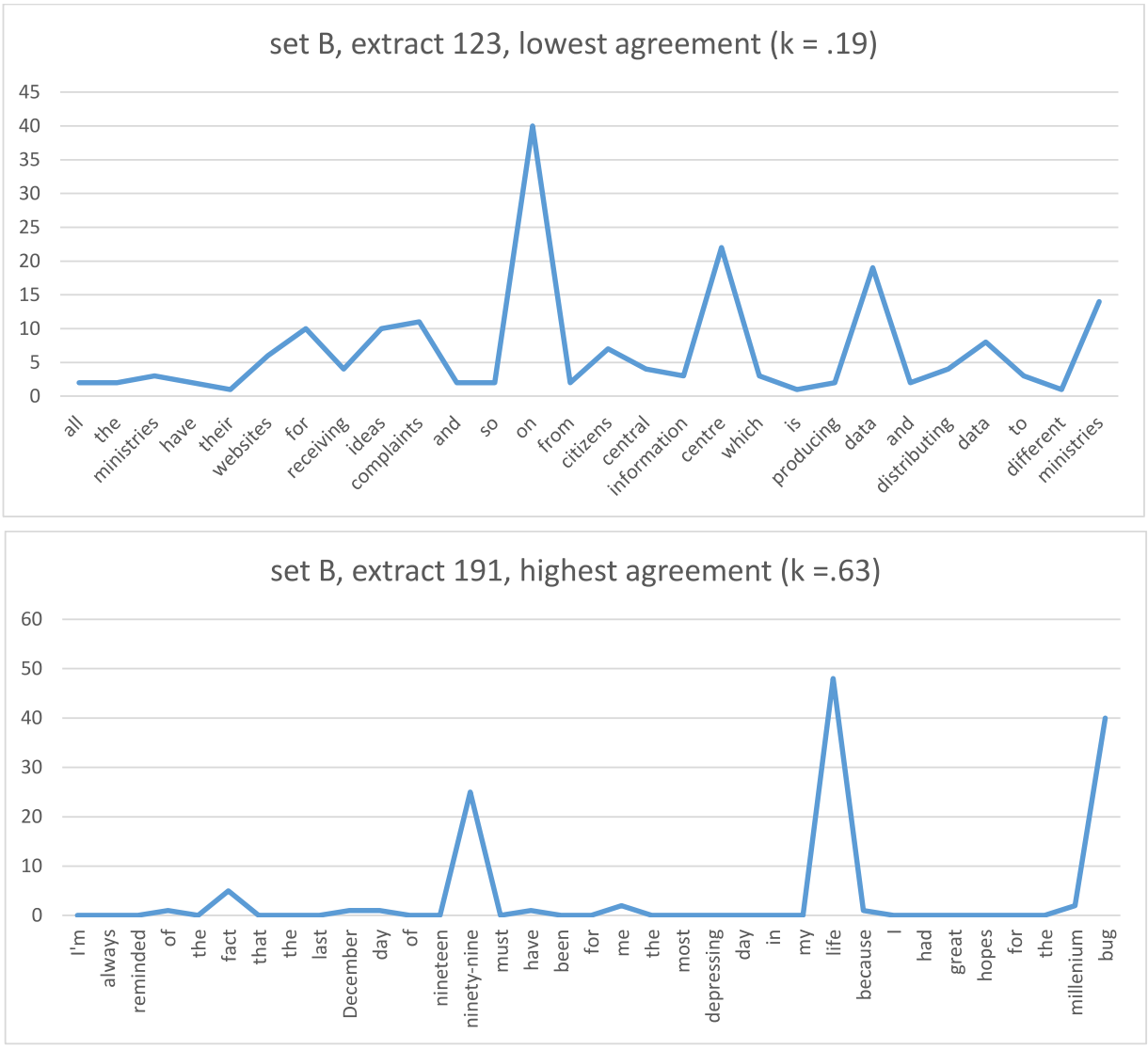


Fig. 4. The extracts with the lowest and the highest kappa in set B.

perspective of statistical significance seems to give a particularly good basis for qualitative analyses since instead of providing an index of overall agreement, as Fleiss' kappa does, it identifies boundaries and non-boundaries which reached 'significant consensus'. For example, boundaries which are marked by as few as 12–15 participants are significant boundaries even though from the perspective of percent agreement they could be treated as non-boundaries. Further, since the ratio between boundary and non-boundary markings seems to stay roughly similar across the participants, the probability distribution of boundary frequencies depends critically on the number of participants and we can conclude that there is a minimum number of participants which is required to ascertain which boundaries are statistically significant non-boundaries (lower tail). Our experiments showed that about 50 participants are enough for this purpose.

Thus, when compared to the analysis of Fleiss' kappa, the analysis of the statistical significance of boundary frequencies provides complementary information: (1) it allows us to assess the statistical significance of boundary frequency after each word, while Fleiss' kappa provides a general measure of agreement per extract or the entire dataset and (2) it allows us to measure consensus on boundaries (upper tail) and non-boundaries (lower tail) separately, while Fleiss' kappa combines this information in one value.

If we now operationalize a chunk as a string of words between statistically significant chunk boundaries, we can compute that the participants identified 730 chunks in set A and 660 chunks in set B. The chunks identified did not recur within either of the sets, apart from a few occurrences of hesitations (*erm, um*) and discourse markers (*okay, first of all, and so on*). In contrast, multi-word units are normally expected to recur. This finding supports the distinction between perceptual and usage-based chunking we suggested in the introduction.

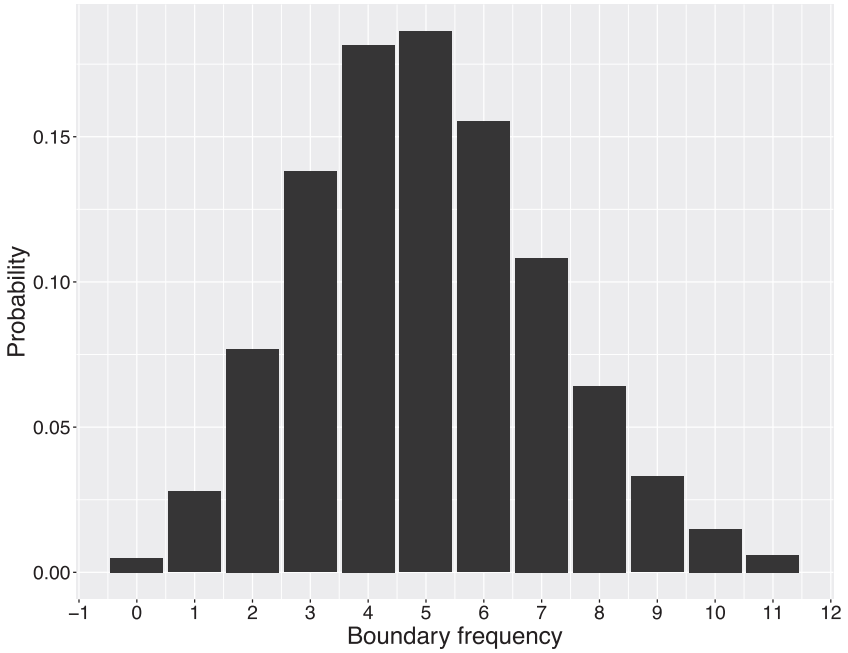


Fig. 5. Null distribution of boundary frequencies for set A.

6. Discussion

This study set out to test the internal validity of chunk identification via crowdsourcing chunk boundary perception data from naïve listeners. Chunking has long been established as a key mechanism in cognition: a domain-general capacity that humans automatically put to use when exposed to a continuous stream of sensory input (e.g. Radvansky & Zacks, 2014; Kurby & Zacks, 2016), and which is likely to play a key role in making sense of language (Christiansen & Chater, 2016). Following Sinclair & Mauranen (2006), we hypothesized that speakers of a language chunk up the speech they hear in roughly similar ways, independently of whether they are L1 or L2 users. We presented the methodology for collecting naïve listener perceptions of chunk boundaries and examined its validity by measuring obtained inter-rater agreement with several methods: percent agreement, Fleiss' kappa, permutation-based null distribution of Fleiss' kappa and permutation-based identification of statistically significant boundary frequencies.

The distribution of boundary frequencies in chunked data appears to approximate Zipfian, like the distribution of many other phenomena in language. This may be taken to reflect the gradient nature of chunk boundaries with an inverse relationship between the relative strength of the boundary and its frequency rank. Most word boundaries in the language are non-boundaries. This property of the data presents challenges for the statistical analysis of inter-rater agreement. While we can clearly distinguish a chunk pattern in participants' responses, which shows that the chunk boundaries are neither randomly nor completely idiosyncratically assigned, capturing this patterning using standard statistical measures is not entirely straightforward.

The observed agreement on boundary placement in our data is remarkable (90–92%), yet high prevalence (.80–.82) results in substantially lower Fleiss' kappa values which range between .43 and .54, indicating only moderate agreement according to the commonly used benchmarks. However, as suggested by many authors, given the impact of prevalence on Fleiss' kappa, the obtained values can only be meaningfully compared in datasets with similar distribution of data across categories. The permutation test we conducted to generate a null distribution of Fleiss' kappa values for our dataset demonstrates that the original Fleiss' kappa value is highly unlikely to have arisen by chance. Further, agreement on non-boundaries as signalled by a very large number of zero boundary frequencies is striking. Permutation tests show that this agreement is statistically significant at .05 level in sets A and B. In other words, it is highly unlikely that all participants will skip a boundary even when the imbalance in marking/non-marking is taken into account. Taken together, these results suggest that naïve listeners strongly agree on chunk boundaries, even if they are L2 speakers of the language. This demonstrates the internal validity of the method and provides support for the construct of perceptual chunking.

The present study establishes a high level of agreement on chunk boundaries among L2 comprehenders. There is no reason to believe that L1 listeners would achieve lower rates of agreement, but this, and many other questions remain open and need to be empirically confirmed in further studies. Another open question is the possible differences between L1 and L2 perceptual chunking. To what degree do L1 and L2 listeners identify the same chunk boundaries in the same data? Do they use or prefer different or similar cues in chunk boundary perception? To what extent can we expect them to weigh cues differently and if so, would different L1 backgrounds predispose L2 users to rely on different weighing strategies? In addition, since this study sought to determine agreement among L2 listeners who understand what they are listening to, we recruited the participants from a pool of fluent L2/ELF speakers

and presented them with speech extracts they were likely to understand. True/false and self-evaluation comprehension questions provided after each extract showed very high comprehension levels: on average, the listeners understood over 90% of the extracts. In future, it will be interesting to examine how the agreement on chunk boundaries changes in less proficient listeners.

Building on the methods developed in this paper, we tested the external validity of crowdsourced chunk boundary perception data in a separate brain imaging experiment. In the experiment, participants listened to the speech extracts from sets A and B while having their brain activity scanned with electro- and magnetoencephalography. Importantly, they did not have access to the transcripts. To find out whether chunk boundaries identified through the chunking task were cognitively real and had a neuronal correlate, we inserted silent pauses in places where listeners' agreement on chunk boundaries or non-boundaries was statistically significant. We found that a pause inserted at a non-boundary, i.e., within a chunk, elicited a biphasic emitted potential which was suggested to reflect a disruption of temporal processing (Besson et al., 1997; Besson & Faïta, 1997). By contrast, at chunk boundaries we observed a Closure Positive Shift (CPS), an event-related potential first discovered in Steinhauer et al. (1999). A CPS is usually elicited at the boundaries of prosodic units, but it was also observed in the absence of prosodic cues (Steinhauer, 2003; Itzhak et al. 2010; Roll et al., 2012; Schremm et al., 2015) and is therefore thought to reflect chunking more generally (Meyer et al. 2020; Henke & Meyer 2021). The observed biphasic emitted potential at non-boundaries and a CPS at boundaries, both identified in a separate chunking experiment, support the external validity of the method. Also, pauses which were placed at boundaries with higher agreement, elicited earlier and more prominent activity. This suggests that chunk boundary strength is gradient and can be measured by the rate of agreement elicited from naïve listeners (for further details see Anurova et al., 2022).

The statistical methods described in this paper can be useful in studying a variety of research questions. For example, a chunk can be defined as a string of words between statistically significant chunk boundaries determined by permutation tests. This operationalization of a chunk allows investigation of chunk properties, such as its average duration. It also provides a good basis for qualitative analysis and linguistic description. It remains to be seen to what extent intuitive chunks correspond to units that can be identified through prosodic or syntactic analysis. As discussed in Section 2, Sinclair & Mauranen (2006) developed a grammatical model which comprehensively describes chunks intuitively identified by two coders. This grammatical model can now be tested on more reliable data.

While the nature of the chunks intuitively identified by naïve listeners was outside the scope of this paper and requires separate analysis, it is clear that the chunks do not correspond to multi-word units. It is therefore important to maintain the distinction between perceptual chunking and usage-based chunking. Through usage-based chunking listeners pick up statistical regularities from the input and in effect learn the language, which can be thought of as an array of such statistical regularities. Perceptual chunking, instead, carves up the input into manageable bits for further processing. These elements do not generally correspond to recurrent multi-word units. If they did, memory constraints would have to be re-thought, since a multi-word unit constitutes one unit by definition, while our memory can process around four according to recent accounts (Cowan, 2001). At the same time, it is likely that learned usage-based chunks play a role in perceptual chunking: at the very least it is clear that a perceptual chunk boundary cannot lie within a multi-word unit, but a more intricate interaction between them is also possible. Given that perceptual chunks do not recur, it is unlikely that they develop into multi-word units with time.

Fleiss' kappa calculated for each extract can serve as convenient index which can be correlated with other extract properties, such as its syntactic features or its comprehensibility. As this study has shown, speech extracts differ in the degree of listener agreement on chunk boundaries. We also found some variation in the rate of agreement across the three datasets we used. In set C, the average rate of agreement was substantially different from sets A and B (Cohen's $d = .593$, $p < .001$). It is likely that the difference arose as a result of the quality of the extracts: for sets A and B, we selected what we judged to be clear and comprehensible extracts, but random extracts for set C. It is reasonable to hypothesise that there is a relationship between inter-rater agreement on chunk boundaries and comprehensibility of the extracts, but this hypothesis calls for a separate study. Similarly, it can be tested to what extent an individual listener's agreement with the group ratings predicts their comprehension of the extracts and language proficiency.

7. Conclusions

Overall, in this study we found support for the internal validity of chunk identification through crowdsourcing naïve listener perception data. The data collected using the chunking task is reliable and can be used in linguistic, applied linguistic and cognitive studies. For example, such data can be used for linguistic description and grammatical analysis. In applied linguistics, it can be used to gauge differences in listening comprehension or, conversely, as an index of extract difficulty, as some extracts are clearly easier to chunk than others. If the relationship between "chunkability" and comprehensibility is confirmed, "chunkability" might be an important focus in developing fluency. For further cognitive exploration, it would seem that there is a wealth of possibilities to search for factors that are involved in chunking and have a bearing on variability in processing. In brief, the methods can be applied to a number of new directions.

Acknowledgments

This study was supported by the Finnish Cultural Foundation (Grant # 00160622). We would like to thank Michal Josífko for programming the first version of *ChunkitApp* and Tuomas Puoliväli for his help in developing the method of testing the statistical significance of boundary frequencies. We are also grateful to the editor of the journal and two anonymous reviewers for their helpful comments on earlier versions of the manuscript.

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. Please note that there may be unintelligible words in the recordings you are going to hear. They appear as '(xx)' in the text.

References

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

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1–L2 differences. *Topics in Cognitive Science, 9*(3), 621–636. 10.1111/tops.12271.

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., & Smith, S. (1982). Functional constraints on sentence processing: A cross-linguistic study. *Cognition, 11*(3), 245–299. 10.1016/0010-0277(82)90017-8.

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.

Besson, M., & Faita, F. (1997). An event-related potential (ERP) study of musical expectancy: Comparison of musicians with nonmusicians. *Journal of Experimental Psychology, 21*, 1278–1296.

Besson, M., Faita, F., Czternasty, C., & Kutas, M. (1997). What's in a pause: Event-related potential analysis of temporal disruptions in written and spoken sentences. *Biological Psychology, 46*(1), 3–23. 10.1016/S0301-0511(96)05215-5.

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

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 115–147. 10.1037/0033-295X.94.2.115.

Bläsing, B. E. (2015). Segmentation of dance movement: Effects of expertise, visual familiarity, motor experience and music. *Frontiers in Psychology, 5*. 10.3389/fpsyg.2014.01500.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology, 10*, 298. 10.3389/fpsyg.2019.00298.

Bornkessel-Schlesewsky, I., Staub, A., & Schlesewsky, M. (2016). The timecourse of sentence processing in the brain. In *Neurobiology of language* (pp. 607–620). Elsevier. 10.1016/B978-0-12-407794-2.00049-3.

Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge University Press.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*(5), 423–429.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 1–52 FirstView. 10.1017/S0140525X1500031X.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. 10.1177/001316446002000104.

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

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114. 10.1017/S0140525X01003922.

Crystal, D. (2012). *English as a global language* (2nd edition). Cambridge University Press.

Dąbrowska, E. (2020). How writing changes language. In A. Mauranen, & S. Vetchinnikova (Eds.), *Language change: The impact of English as a lingua franca* (pp. 75–94). Cambridge University Press.

de Ruiter, J. P., Mitterer, Holger., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language, 82*(3), 515–535. 10.1353/lan.2006.0130.

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. 10.1038/nn.4186.

ELFA (2008). The Corpus of English as a lingua franca in academic settings. Director: Anna Mauranen. http://www.helsinki.fi/elfa

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. 10.1017/9781316091746.006.

Ellis, N. C., Römer, U., & O'Donnell, M. B (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Wiley.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549. 10.1016/0895-4356(90)90158-L.

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

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

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C. H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*(6), 236–243. 10.1016/S1364-6613(00)01662-4.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Goldberg, A. E., Hoffmann, T., & Trousdale, G. (2013). Constructionist approaches. *The oxford handbook of construction grammar*. Oxford University Press. 10.1093/oxfordhb/9780195396683.013.0002.

Grabowski, K. C., & Oh, S. (2018). Reliability analysis of instruments and data coding. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 541–565). Palgrave Macmillan UK. 10.1057/978-1-137-59900-1_24.

Gries, S. T., & Wulff, S. (2005). Do foreign language learners also have constructions? *Annual Review of Cognitive Linguistics, 3*(1), 182–200.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical & Statistical Psychology, 61*(1), 29–48. 10.1348/000711006X126600.

Gwet, K. L. (2014). *Handbook of inter-rater reliability, 4th edition: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34. 10.20982/tqmp.08.1.p023.

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

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

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science, 44*(7), e12848. 10.1111/cogs.12848.

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), 8–13. 10.1097/WNR.0b013e328330251d.

Jenkins, J. (2015). *Global englishes: A resource book for students*. Routledge.
Kaltenböck, G., Heine, B., & Kuteva, T. (2011). On thetical grammar. *Studies in Language, 35*(4), 852–897. 10.1075/sl.35.4.03kal.
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. 10.1523/JNEUROSCI.0302-20.2020.
Kidd, E., Arciuli, J., Christiansen, M. H., Isbilen, E. S., Revius, K., & Smithson, M. (2020). Measuring children's auditory statistical learning via serial recall. *Journal of Experimental Child Psychology, 200*, Article 104964. 10.1016/j.jecp.2020.104964.
Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences, 12*(2), 72–79. 10.1016/j.tics.2007.11.004.
Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363–374. 10.2307/2529786.
Larsson, T., Paquot, M., & Plonsky, L. (2020). Inter-rater reliability in learner corpus research: Insights from a collaborative study on adverb placement. *International Journal of Learner Corpus Research, 6*(2), 237–251. 10.1075/ijlcr.20001.lar.
MacWhinney, B. (2000). *The childes project: Tools for analyzing talk, volume II: The database*. Psychology Press. 10.4324/9781315805641.
MacWhinney, B., Kail, M., & Hickmann, M. (2012). A tale of two paradigms. In *Language acquisition and language disorders: 52* (pp. 17–32). John Benjamins. 10.1075/lald.52.03mac.
MacWhinney, B., Bates, E., & Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior, 23*(2), 127–150. 10.1016/S0022-5371(84)90093-8.
Mahrt, T. (2016). LMEDS: Language markup and experimental design software. URL https://github.com/timmahrt/LMEDS.
Mauranen, A. (2012). *Exploring ELF: Academic English shaped by non-native speakers*. Cambridge University Press.
McCauley, S. M., & Christiansen, M. H. (2014). A computational model. *Mental Lexicon, 9*(3), 419–436. 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. 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* cercor;bhw228v1. 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. 10.1080/23273798.2020.1827155.
Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology, 28*(1), 28–38. 10.1037/h0035584.
Pawley, A., & Syder, F. H (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards, & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). London: Longman.
Perruchet, P. (2005). Statistical approaches to language acquisition and the self-organizing consciousness: A reversal of perspective. *Psychological Research Psychologische Forschung, 69*(5–6), 316–329. 10.1007/s00426-004-0205-6.
Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*(2), 246–263. 10.1006/jmla.1998.2576.
Perruchet, P., & Vinter, A. (2002). The self-organizing consciousness. *Behavioral and Brain Sciences, 25*(3), 297–330. 10.1017/S0140525X02000067.
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). 10.2202/1544-6115.1585.
Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130. 10.3758/s13423-014-0585-6.
Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538–553. 10.1111/modl.12335.
Plonsky, L., & Oswald, F. L. (2014). How big is "Big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. 10.1111/lang.12079.
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*, Article 108654. 10.1016/j.jneumeth.2020.108654.
Radvansky, G. A., & Zacks, J. M. (2014). *Event Cognition*. Oxford University Press.
Rimmele, J. M., Poeppel, D., & Ghitza, O. (2020). Acoustically driven cortical delta oscillations underpin perceptual chunking bioRxiv. 2020.05.16.099432 10.1101/2020.05.16.099432
Roll, M., Lindgren, M., Alter, K., & Horne, M. (2012). Time-driven effects on parsing during reading. *Brain and Language, 121*(3), 267–272. 10.1016/j.bandl.2012.03.002.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 74*(5294), 1926–1928.
Schremm, A., Horne, M., & Roll, M. (2015). Brain responses to syntax constrained by time-driven implicit prosodic phrases. *Journal of Neurolinguistics, 35*, 68–84. 10.1016/j.jneuroling.2015.03.002.
Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193–247. 10.1007/BF01708572.
Simpson, R. A., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). *The Michigan corpus of academic spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.
Sinclair, J. McH (1991). *Corpus, concordance, collocation*. Oxford University Press.
Sinclair, J. McH, & Mauranen, A (2006). *Linear unit grammar integrating speech and writing*. John Benjamins.
Speer, N. K., Swallow, K. M., & Zacks, J. M. (2003). Activation of human motion processing areas during event perception. *Cognitive, Affective, & Behavioral Neuroscience, 3*(4), 335–345. 10.3758/CABN.3.4.335.
Sridharan, D., Levitin, D. J., Chafe, C. H., Berger, J., & Menon, V. (2007). Neural dynamics of event segmentation in music: Converging evidence for dissociable ventral and dorsal networks. *Neuron, 55*(3), 521–532. 10.1016/j.neuron.2007.07.003.
Steinhauer, K. (2003). Electrophysiological correlates of prosody and punctuation. *Brain and Language, 86*(1), 142–164. 10.1016/S0093-934X(02)00542-4.
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. 10.1038/5757.
Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation, 9*, 1–12. 10.7275/96JP-XZ07.
Vetchinnikova, S. (2019). *Phraseology and the advanced language learner*. Cambridge University Press. 10.1017/9781108758703.
Vetchinnikova, S., Mauranen, A., & Mikušová, N. (2017). ChunkitApp: Investigating the relevant units of online speech processing. In *Proceedings of INTERSPEECH 2017 –18th annual conference of the international speech communication association* (pp. 811–812).
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.
Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes, 25*(7–9), 905–945. 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. 10.1080/01690960444000070.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press. 10.1017/CBO9780511519772.
Yorio, C. A. (1989). Idiomaticity as an indicator of second language proficiency. In K. Hyltenstam, & L. K. Obler (Eds.), *Bilingualism across the lifespan: Aspects of acquisition, maturity and loss* (pp. 55–72). Cambridge University Press.