“Um…, it’s really difficult to... um… speak fluently”: Neural tracking of spontaneous speech

Galit Agmon[1,2], Manuela Jaeger[3], Reut Tsarfaty[4],

Martin G Bleichner*[3], Elana Zion Golumbic*[1]

[1] The Gonda Center for Multidisciplinary Brain Research, Bar Ilan University, Ramat Gan, Israel

[2] Frontotemporal Degeneration Center, Perelman School of Medicine, University of Pennsylvania, USA

[3] Neurophysiology of Everyday Life Group · Department of Psychology, University of Oldenburg, Oldenburg Germany

[4]Dept of Computer Science, Bar Ilan University, Ramat Gan, Israel

* joint senior authorship

Abstract

Spontaneous real-life speech is imperfect in many ways. It contains disfluencies and ill-formed utterances that the brain needs to contend with in order to extract meaning out of speech. Here, we studied how the neural speech-tracking response is affected by three specific factors that are prevalent in spontaneous colloquial speech: (1) the presence of non-lexical fillers, (2) the need to detect syntactic boundaries in disfluent speech and (3) the effort involved in processing syntactically complex phrases. Neural activity (EEG) was recorded from individuals as they listened to an unscripted, spontaneous narrative, which was analyzed in a time-resolved fashion to identify fillers, detect syntactic boundaries and assess the syntactic complexity of different phrases. When considering these factors in the speech-tracking analysis, we found that it was affected by all of them. The most consistent effect, observed for all three factors, was modulation of a centro-frontal negative response that peaked around 350 ms, highly resembling the well-known N400 ERP response linked to various aspects of lexical access and semantic processing. This response was observed for lexical words but not for fillers, was larger for opening vs. closing words of a clause and was enhanced in response to high-complexity phrases. These findings broaden ongoing efforts to understand neural processing of speech under increasingly realistic conditions. They highlight the importance of considering the imperfect nature of real-life spoken language, linking past research on linguistically well-formed and meticulously controlled speech to the type of speech that the brain actually deals with on a daily basis.

Introduction

Neural speech tracking has become an increasingly useful tool for studying how the brain encodes and processes continuous speech (Obleser and Kayser, 2019; Brodbeck and Simon, 2020). Importantly, characterizing linguistic attributes of speech on a continuous basis gives a new angle to auditory attention and neurolinguistics, as researchers have been able to dissociate neural responses driven by the acoustics of speech from those capturing higher-order processes in a dynamically changing speech signal (Brodbeck et al., 2018; Keitel et al., 2018; Inbar et al., 2020; Gillis et al., 2021). And yet, the speech stimuli used in these studies are generally highly scripted and edited, taken – for example – from audiobooks or TED talks, which are also extensively rehearsed and are delivered by professionals. These stimuli are in many ways different from colloquial speech that we hear every day (Mehta and Cutler, 1988; Blaauw, 1994; Haselow, 2017). Spontaneous speech is markedly less fluent than scripted speech and often includes partial or grammatically incorrect syntactic structures (Shriberg, 2001; Haselow, 2017 p.1). These disfluencies may render spoken speech less trivial to parse compared to the planned well-structured sentences in scripted speech materials (see, for example, Figure 1A). By focusing mostly on scripted materials, we likely underestimate important higher-order speech processes and therefore the generalizability of the findings to real-life speech is limited. Addressing this gap, in the current EEG study, we assess neural speech tracking of spontaneously generated speech. We focus on two independent challenges the brain has to cope with when processing spontaneous speech. The first challenge is caused by the abundant presence of disfluencies and fillers. The second challenge relates to the online segmentation and syntactic analysis of the speech stream. We will discuss these challenges in the following few sections.

Disfluency and fillers

A prominent feature of spontaneous speech is that it contains frequent pauses, self-corrections and repetitions (Fox Tree, 1995; Clark and Wasow, 1998; Bortfeld et al., 2001). These disfluencies are generally accompanied by fillers, which are non-lexical utterances (or filled-pauses) such as "um" or "uh", or discourse markers such as "you know" and "I mean" (Fox Tree and Schrock, 2002; Tottie, 2014). Fillers can take on different forms, however, their prevalence across a multitude of spoken languages (e.g. Wieling et al., 2016; Tian et al., 2017) suggests it is a core feature of spontaneous speech. Although fillers do not, in and of themselves, contribute specific lexical information, they are also not mere glitches in speech production. Rather, they likely serve several important communicative goals, helping the speaker in transforming their internal thoughts to speech and helping the listener interpret this speech. Specific roles that have been attributed to fillers include signaling hesitation in speech planning (Corley and Stewart, 2008), conveying lack of certainty in the content (Smith and Clark, 1993; Brennan and Williams, 1995), serving as a cue to focus attention on upcoming words or complex syntactic phrases (Fox Tree, 2001; Clark and Fox Tree, 2002; Watanabe et al., 2008; Fraundorf and Watson, 2011), signaling unexpected information to come (Arnold et al., 2004; Corley et al., 2007; Barr and Seyfeddinipur, 2010) and disambiguating syntactic structures (Bailey and Ferreira, 2003). Moreover, studies have shown that the presence of fillers improves accuracy and memory of speech content (Brennan and Schober, 2001; Fox Tree, 2001; Corley et al., 2007; Fraundorf and Watson, 2011) and that surprise-related neural responses (event-related potentials; ERPs) to target words are reduced if they are preceded by a filler (Corley et al., 2007; Collard et al., 2008). And yet, despite their clearly important role in the production and perception of spontaneous speech, fillers and other disfluencies are generally absent in planned speech that is used in most lab experiments. Therefore, how the brain processes fillers has not been studied extensively.

Syntactic analysis

The disfluent nature of spontaneous speech increases its syntactic complexity and leads to syntactic irregularities. Unlike written text or highly edited spoken scripts, spontaneous speech does not contain clear sentence endings, nor is it always grammatically organized. Rather, sentences and utterances are constructed ‘on the fly’, sometimes yielding extremely long and complex sentences, as well as syntactically incomplete utterances (e.g. *“and, so, then we went to the bus, but it, like, didn’t come, the bus back home I mean, so we had to wait, I really don't know for how long”*). This complexity poses two distinct challenges to the listener: how to parse the continuous input-stream into meaningful syntactic units correctly, and how to derive meaning from these complex syntactic units. Therefore, we will focus on syntactic parsing and complexity and investigate how the brain deals with these challenges in processing spontaneous speech.

Syntactic parsing: Syntactic parsing is critical for speech comprehension, even under assumptions of a "good-enough" or noisy parse (Ferreira and Patson, 2007; Traxler, 2014). To detect syntactic boundaries in spoken language, listeners rely on their accumulated syntactic analysis of an utterance as well as on prosodic cues such as pauses and changes in pitch and duration (Fodor and Bever, 1965; Garrett et al., 1966; Strangert, 1992; Strangert and Strangert, 1993; Langus et al., 2012; Hawthorne and Gerken, 2014; Ding et al., 2016; Kaufeld et al., 2020; Har-shai Yahav and Zion Golumbic, 2021). There are some behavioral and neural indications that words occurring at the final position of syntactic structures have a ‘special status’. In production, words in final positions tend to be prosodically marked (Klatt, 1975; Cooper and Paccia-Cooper, 1980). In comprehension, reading studies show that sentence-final words have prolonged reading and fixation times as well as increased ERP responses. These effects are known as ‘wrap up effects’, which is a general name for the integrative processes triggered by the final word (Just and Carpenter, 1980; Stowe et al., 2018 for review). Independently, there is also evidence for neural responses associated with

detecting prosodic breaks (Steinhauer et al., 1999; Pannekamp et al., 2005; Peter et al., 2014), which often coincide with syntactic boundaries. However, the neural correlates of syntactic boundaries have seldom been studied in the context of spoken language, and particularly not for spontaneous speech, where sentence boundaries are not as well-formed as in scripted language materials. Here, we will study the neural response to the first and last word of major syntactic units.

Syntactic complexity: Syntactic complexity has been traditionally studied through designed contrasts. One of the most studied cases of syntactic complexity is the well-known contrast between object-relative-clauses and subject-relative-clauses, which we will demonstrate through the following example: The following two sentences use the same set of words and convey similar meanings: *"The woman <u>that John hugged</u> is Yoko" and "The woman <u>that hugged John</u> is Yoko"*. However, the former sentence (object-relative-clause) is more syntactically complex than the latter (subject-relative-clause), as attested by a plethora of studies on acquisition and processing (for a recent review, see Lau and Tanaka, 2021). Such contrasts in syntactic complexity have been reported to affect primarily late neural responses such as the N400 and the late anterior negativity (Kluender and Kutas, 1993; King and Kutas, 1995; Müller et al., 1997; Fiebach et al., 2001, 2002; Friederici et al., 2002; Phillips et al., 2005; Ueno and Garnsey, 2008; Gouvea et al., 2010; Kaan et al., 2010; Kwon et al., 2013; Bulut et al., 2018). Despite the important insights that designed-contrasts studies provide on syntactic processing, it is difficult to implement such methods on uncontrolled sentences in spontaneous speech. Building on those experiments, we used novel methods to test whether similar effects of syntactic complexity are observed in the neural tracking of spontaneous speech. Sentences in which syntactically related words are far apart from each other are considered to be more complex than sentences where the syntax follows the linear word-order more closely (notice that the object *woman* in the more complex sentence is farther

apart from its related verb *hugged*, compared to the less complex sentence). This is because long-distance relationships between words require listeners to maintain information in working memory for a longer time before it can be integrated into appropriate syntactic structures (Caplan et al., 1998; Gibson, 1998; Ben-Shachar et al., 2003; Grodzinsky and Santi, 2008; Kaan et al., 2010). We used a metric that quantifies the long-distance relationships in a sentence, so we could calculate syntactic complexity for spontaneous, uncontrolled, speech.

To summarize, spontaneous speech differs in many ways from scripted speech. Here we focused on three main characteristics of spontaneous speech that listeners need to deal with when listening to another person: first, the presence of fillers; second, the detection of syntactic boundaries; and third, the effect of syntactic complexity. Our stimulus consisted of a six-minute long monologue rich in those features, recorded from a speaker spontaneously recounting a personal experience. We analyzed the monologue to identify fillers, detect syntactic boundaries and calculate syntactic complexity. These served as data-driven regressors for analyzing the neural activity measured from participants listening to the monologue. In doing so, this study aims to bridge the gap between the vast literature studying brain responses to meticulously controlled speech and language materials, and the type of speech materials that the brain deals with on a regular daily basis.

Methods

Participants

Twenty participants took part in this experiment. All participants were native Hebrew speakers, right-handed (16:4 F:M; age range 20-30, average 23.1±2.65). Prior to their participation, participants signed informed consent approved by the IRB of Bar-Ilan University and were compensated for their participation with credit or payment.

Stimuli and Procedure

The stimulus was a single 6-minute long recording of a personal narrative, told in the first person by a female native Hebrew speaker. The narrative was neutral, unscripted, and spontaneously generated, and described the speaker's participation in a Facebook group called "stories with no point" and a social face-to-face meet-up organized by members of the group. The only instructions given to participants were to listen passively to the story presented without breaks. This 6-minute session was used as an interlude between two parts of another experiment (Har-shai Yahav et al., in progress), that focused on changes in auditory responses over time and is orthogonal in its goals to the data reported here.

For the purpose of linguistic analysis, the narrative was transcribed manually and the onset and duration of each word were identified and time-stamped by two independent annotators, and confirmed by a third using the software Praat (Boersma and Weenink, 2021). As detailed below, this time-stamped transcript was used to detect the presence of fillers, identify syntactic boundaries, and calculate phrases' syntactic complexity.

EEG recordings

EEG was recorded using a 64 Active-Two system (BioSemi) with Ag-AgCl electrodes, placed according to the 10-20 system, at a sampling rate of 1024 Hz. Additional external electrodes were used to record from the mastoids bilaterally and both vertical and horizontal EOG electrodes were used to monitor eye movements. The experiment was conducted in a dimly lit, acoustically, and electrically shielded booth. Participants were seated on a comfortable chair and were instructed to keep as still as possible and breathe and blink naturally. Experiments were programmed and presented to participants using PsychoPy (https://www.psychopy.org) (Peirce et al., 2019).

EEG preprocessing and speech-tracking analysis

EEG preprocessing and analysis were performed using the matlab-based FieldTrip toolbox (Oostenveld et al., 2011) as well as custom-written scripts. Raw data was first visually inspected, and time-points with gross artifacts exceeding ±50 μV (and were not eye movements) were removed. Independent Component Analysis (ICA) was performed to identify and remove components associated with horizontal or vertical eye movements as well as heartbeats. Any remaining noisy electrodes that exhibited either extreme high-frequency activity (>40Hz) or low-frequency activity/drifts (<1 Hz), were replaced with the weighted average of their neighbors using an interpolation procedure. The clean EEG data was filtered between 1-10Hz. The broadband envelope of the speech was extracted using an equally spaced filterbank between 100 to 10000 Hz based on Liberman's cochlear frequency map (Liberman, 1982). The narrowband filtered signals were summed across bands after taking the absolute value of the Hilbert-transform for each one, resulting in a broadband envelope signal. The speech-envelope and EEG data were aligned in time and downsampled to 100Hz, for computational efficiency.

Speech-tracking analysis was performed using scripts from the STRFpak Matlab toolbox (strfpak.berkeley.edu) which were adapted for EEG data (Zion Golumbic, Cogan, et al. 2013, Har-shai Yahav 2021). In this approach, a normalized reverse correlation is applied in order to estimate a Temporal Response Functions (TRF) that best describes the linear relationship between features of the presented speech stimulus S(t) and the recorded neural response R(t). Speech tracking analysis was performed separately for each of the four-stimulus features of interest as described below:

(1) Full Envelope Acoustic Model: The full envelope acoustics model was estimated using the entire stimulus and the corresponding EEG data, which was segmented into ten equal-length non-overlapping epochs (36 seconds each). To minimize the effects of over-fitting, we used a jackknife leave-one-out cross-validation procedure on 90% of the data to select

sparseness and tolerance values for the TRF model that explain the largest amount of variance in the EEG response. We then tested the resulting TRF on the remaining 10% of the data (not included in the training) to obtain a measure of its predictive power, which is the Pearson's correlation (r) between the estimated and actual EEG signal, at each electrode. This train-test procedure was repeated five times, using different train-test selections, and the average TRF and predictive power are reported. The sparseness and tolerance values selected for the full-envelope model were then kept constant for the estimation of TRFs in all subsequent analyses.

(2) Lexical words vs. Fillers: Fillers are sounds or utterances that do not carry any specific linguistic information, but nonetheless punctuate spontaneous speech. These can include filled-pauses (e.g., "um", "uh") and filler discourse markers (e.g., "like") (Fox Tree and Schrock, 2002; Tottie, 2014). In the current study, fillers were identified and marked as part of the transcription process to examine how they are encoded in the neural response to spontaneous speech. We used the time-stamped transcription to identify where fillers occurred in the speech narrative to evaluate the neural response to fillers embedded within real-life speech. We then separated the full acoustic envelope into two regressors – one containing only the envelope of the fillers and the other containing only the lexical portions of the narrative (without the fillers). Since fillers make up only 18.5% of the entire speech stimulus (94 fillers vs. 561 words), direct comparison between the full-lexical regressor and the filler-regressor may be biased by the different proportion of the stimulus that each one represents. To control for this, we created a proportion-controlled lexical regressor, which consisted of a subset of 94 non-filler words and thus was more similar to the filler regressor (Figure 2A). TRFs were estimated for these two regressors and were compared statistically, using a paired t-test at each time-point and at each electrode, and corrected for multiple comparisons using threshold-free cluster correction (TFCE; implemented in fieldtrip).

(3) Syntactic Boundaries (Opening vs. Closing words): To study the detection of syntactic boundaries, we first needed to identify these boundaries in the speech narrative presented to participants. For this analysis, we marked the boundaries of various types of syntactic units, referred to as "clauses". We define the main clause as the minimal unit containing a predicate and all its complements and modifiers. In cases of subordination, complements can be considered clauses of their own (e.g., the complement clause is bracketed in "we decided [that we would meet in one of the gardens]"). We also marked the boundaries of adverbial clauses (e.g., the bracketed part in "we met [because we needed to talk]"), as well as the boundaries of coordinate clauses (connected with "and" or "but"), relative clauses, incomplete utterances, and heavy noun phrases.

Our marking of syntactic boundaries was done primarily based on the transcript of the speech. However, it was double-checked relative to the audio-recording in search of cases where the spoken prosody suggested a different intended parsing than the textual syntactic analysis. For example, the text "I went home with my friends" could be considered to be a single clause from a purely textual perspective. However, in the audio recording used here, the speaker inserted a pause after the word “home”, and hence, a listener would likely have identified the word "home" as the final word in the clause, before the speaker decided to continue it with a prepositional phrase ("with my friends").

Based on this manual segmentation, we identified the individual words serving as opening and closing-words of clauses to compare their respective neural responses. We then created two separate regressors capturing only the parts of the envelope including opening or closing words, respectively (Figure 3A). TRFs were estimated for these two regressors and were compared statistically, using a paired t-test at each time-point and at each electrode, and corrected for multiple comparisons using threshold-free cluster correction (TFCE; implemented in fieldtrip).

(4) Syntactic Complexity: Our final goal was to assess the effects of syntactic complexity on the neural encoding of spontaneous speech. To this end, we calculated the syntactic complexity of each main clause (defined, as described above, as a predicate with all its complements, including complement clauses), adverbial clauses, non-restrictive relative clauses, and fragments. The syntactic complexity of each clause was quantified using a measure of mean dependency distance (MDD): the dependency distance (DD) between two syntactically related words is the number of words that separates them. For example, in the sentence “I wrote a long email”, there is a DD of 1 between the verb ("wrote") and its subject ("I"), but there is a DD of 3 between the verb ("wrote") and its object ("email"). The mean dependency distance (MDD) of an entire clause is taken as the average of all the dependencies it contains (Hudson, 1995). To calculate the MDD, the dependency structure of each clause was analyzed using an automatic parser for Hebrew (More et al., 2019), and manually corrected by an expert linguist (GA). Then, each word was given a DD score reflecting the distance from its head, which was averaged across all the words in the utterance to produce the MDD (see Figure 1B). Initial words that serve as connectors between clauses (e.g., "and", "but", "because") were given a DD score of 1.

To compare how the speech tracking response was affected by MDD, we segmented the EEG data into smaller segments corresponding to each individual clause (Figure 5B&C top) and separated clauses with high vs. low MDD, using a median split. However, we noticed that this measure was positively correlated with clause-length (Pearson’s correlation r = 0.58; Figure 5A), since higher complexity is only possible in longer clauses. To avoid a potential confound between length and complexity, we split all clauses into two groups of long and short clauses, such that the cumulative duration of the clauses in each group represented roughly half of the full narrative (vertical line in figure 5A). TRFs were estimated for clauses with high and low MDD (separately for short and long clauses), and were compared

statistically, using a paired t-test at each time-point and each electrode, and corrected for multiple comparisons using threshold-free cluster correction (TFCE; implemented in fieldtrip).

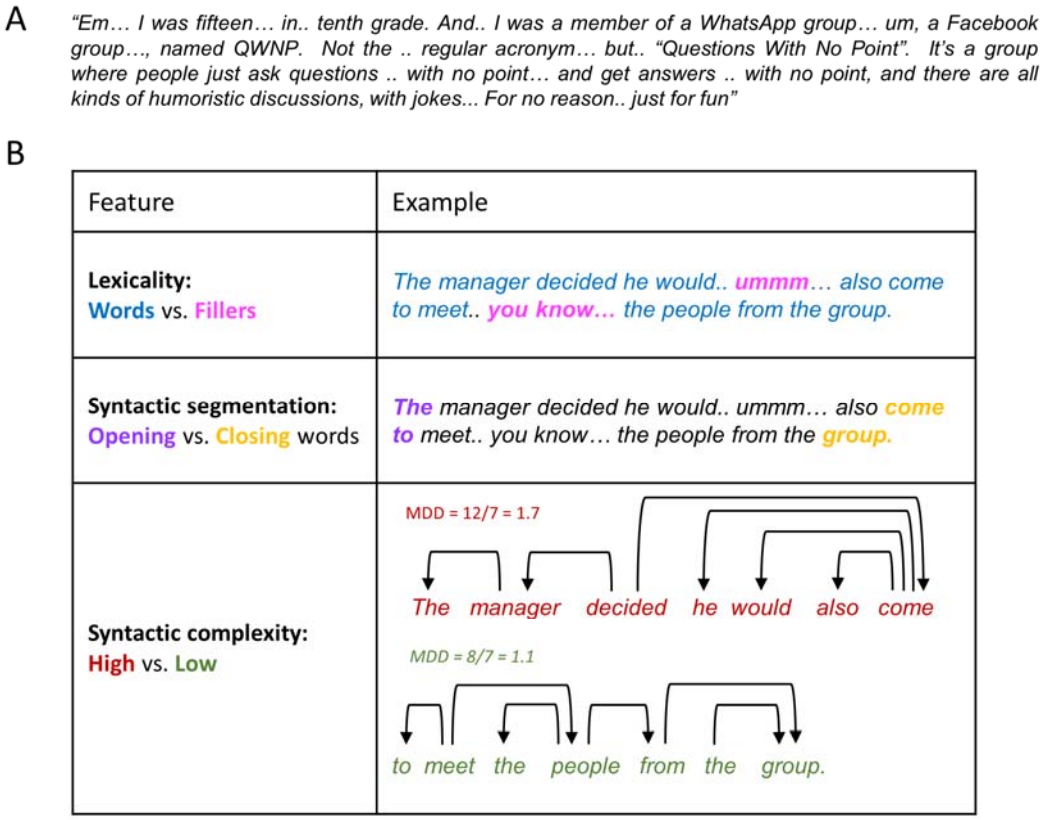

Figure 1. A. Excerpt from the speech stimulus used, demonstrating the disfluencies of spontaneous speech. **B.** Illustration of the three features of spontaneous speech analyzed here. *Top*: Color coded distinction between lexical words that make up the content of the speech (blue) and fillers (magenta; including both filled-pauses and filler-words). *Middle*: Color coded distinction between opening (purple) and closing words (yellow) of clauses. *Bottom:* Demonstration of the calculation of mean dependency distance (MDD) for syntactic complexity, shown for two sentences with high (red) and low (green) MDD values. Arcs represent dependency relationship between words in the sentence, and the MDD is the average of distances across all words (main verb "meet", which has no head, is defined as having a distance of 0).

Results

Full acoustic model

Speech-tracking analysis using the full acoustic envelope yielded a robust TRF with four prominent peaks at 50 ms, 100 ms, 200 ms, and 300-400 ms, in line with the morphology reported in previous studies with the best predictive power values in fronto-central electrodes (Figure 2). This analysis is shown primarily for methodological validation purposes, showing the good predictive power of the model and the time-course of the estimated TRFs. We now present how the different features of spontaneous speech modulated this response.

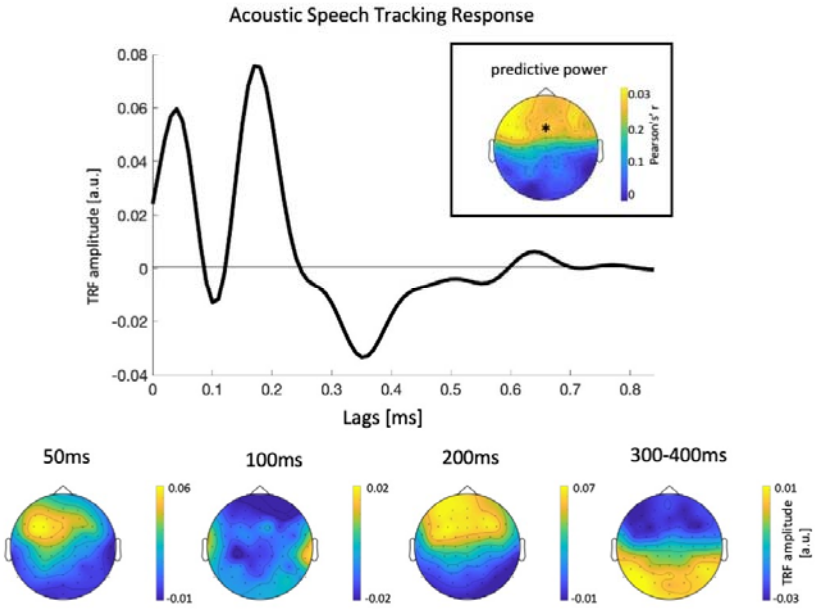

Figure 2. *Top*: The speech-tracking TRF derived using the full speech-envelope (Acoustic model), shown from the centro-frontal electrode Fz (location indicated with an asterisk in the inset). *Inset*: Topographical distribution of the predictive-power of the model. *Bottom*: Topographical distributions of the TRF amplitude at the four main peaks in the waveform: at 50ms, 100ms, 200ms and the average between 300-400 ms.

Neural response to fillers

The TRF estimated using only the proportion-controlled lexical regressor was qualitatively similar to the acoustic-TRF, with all four-peaks dominantly present albeit some of them were smaller, likely due to the substantially reduced proportion of the stimulus it represents. The TRF for the filler regressor shares some of those peaks, particularly in the early time windows, however, the response was substantially reduced relative to the proportion-controlled lexical regressors. Significant differences were observed in the time window between 50-150 ms and between 300-400ms ($p<0.05$, TFCE cluster corrected; Figure 3).

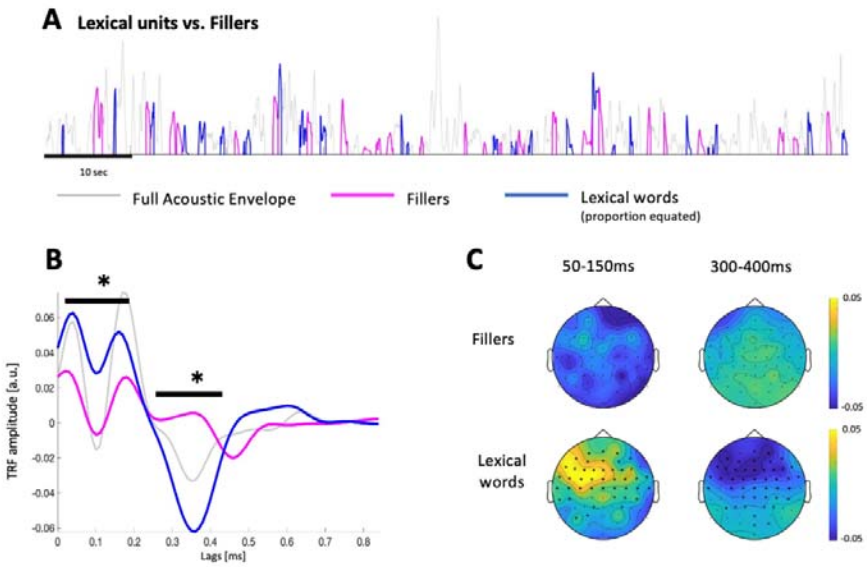

Figure 3. A. A portion of the acoustic envelope (gray), overlayed with the regressor capturing the fillers (magenta) and proportion-equated lexical regressors (blue). **B.** The TRF derived for the filler (blue) and lexical (magenta) regressors. The TRF for the full acoustic model is shown in grey, for references (same as in Figure 2). Times with significant differences between the TRFs are indicated with vertical lines and an asterisk ($p<0.05$ TCFE cluster corrected). **C.** Scalp topographic of the TRF response in time-windows where significantly larger responses were found between for the lexical regressor vs. fillers. Electrodes where the effect was significant are indicated by asterisks on the lower topographies.

Syntactic Boundaries

We next tested whether the neural tracking responses show evidence for the detection of syntactic boundaries. To this end, we analyzed the time-stamped to identify the boundaries between syntactic utterances and marked the position of the opening and closing words in each utterance. Two regressors were created containing the acoustic envelopes of opening vs. closing words (Figure 4) and TRFs were estimated separately for each regressor. Significant differences in the neural responses to opening vs. closing words were found in three time-windows (all $p<0.05$, TFCE cluster corrected): Opening words elicited a larger early positive response between 30-80 ms and a larger negative response around 200-300 ms, whereas responses to closing words elicited a larger positive response at 200 ms.

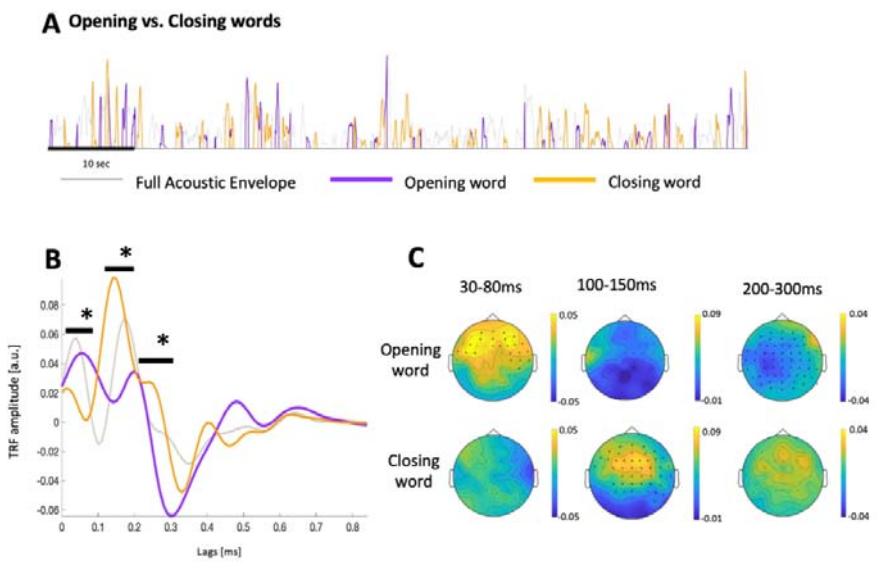

Figure 4. A portion of the acoustic envelope (gray), overlayed with the regressor capturing the opening words (purple) and closing words (yellow) of clauses. **B.** The TRF derived for the opening word (purple) and closing word (yellow) regressors. The TRF for the full acoustic model is shown in grey, for references (same as in Figure 2). Times with significant differences between the TRFs are indicated with vertical lines and an asterisk ($p<0.05$ TCFE cluster corrected). **C.** Scalp topographic of the TRF response in time-windows where significantly larger responses were found between opening and closing words. Electrodes where the effect was significant are indicated by asterisks on topography where the TRF response was highest.

Syntactic Complexity

Last, we were interested in assessing whether the neural speech tracking response was affected by the syntactic complexity, quantified as the mean dependency distance (MDD) of each clause. However, we noticed that this measure was highly positively correlated with clause length (Pearson's correlation r = 0.58; Figure 5A), since shorter clauses are less likely to contain complex syntax. To avoid a potential confound between length and complexity, we analyzed separately long and short clauses, defined such that the cumulative duration of the clauses in each group represented roughly half of the full narrative (threshold = 3.5 sec; vertical line in figure 5A). We then tested the effects of complexity separately for short and long clauses, using a median split of MDD values (median MDD = 1 for short clauses and median MDD = 1.7 for long clauses). Visual inspection of the resulting TRFs indicated that the timing of the TRF components was slightly shifted for the short vs. long clauses, with longer latency peaks found for the short vs. long clauses (Figure 5B&C middle). Even though this was not an immediate goal of the current study, this result justified our choice to perform separate analyses for short and long clauses. Importantly, for both short and long clauses, the TRF responses were significantly higher for clauses with high vs. low MDD (Figure 5B&C middle; all $p<0.05$ TFCE cluster corrected). For the short clauses, MDD was found to affect an early positivity between (100-140ms) and a later negativity between 350-450ms. For long clauses, we also found an increased negative response in a similar time window of 270-340 ms, as well as an increased later positive response between 360-410 ms.

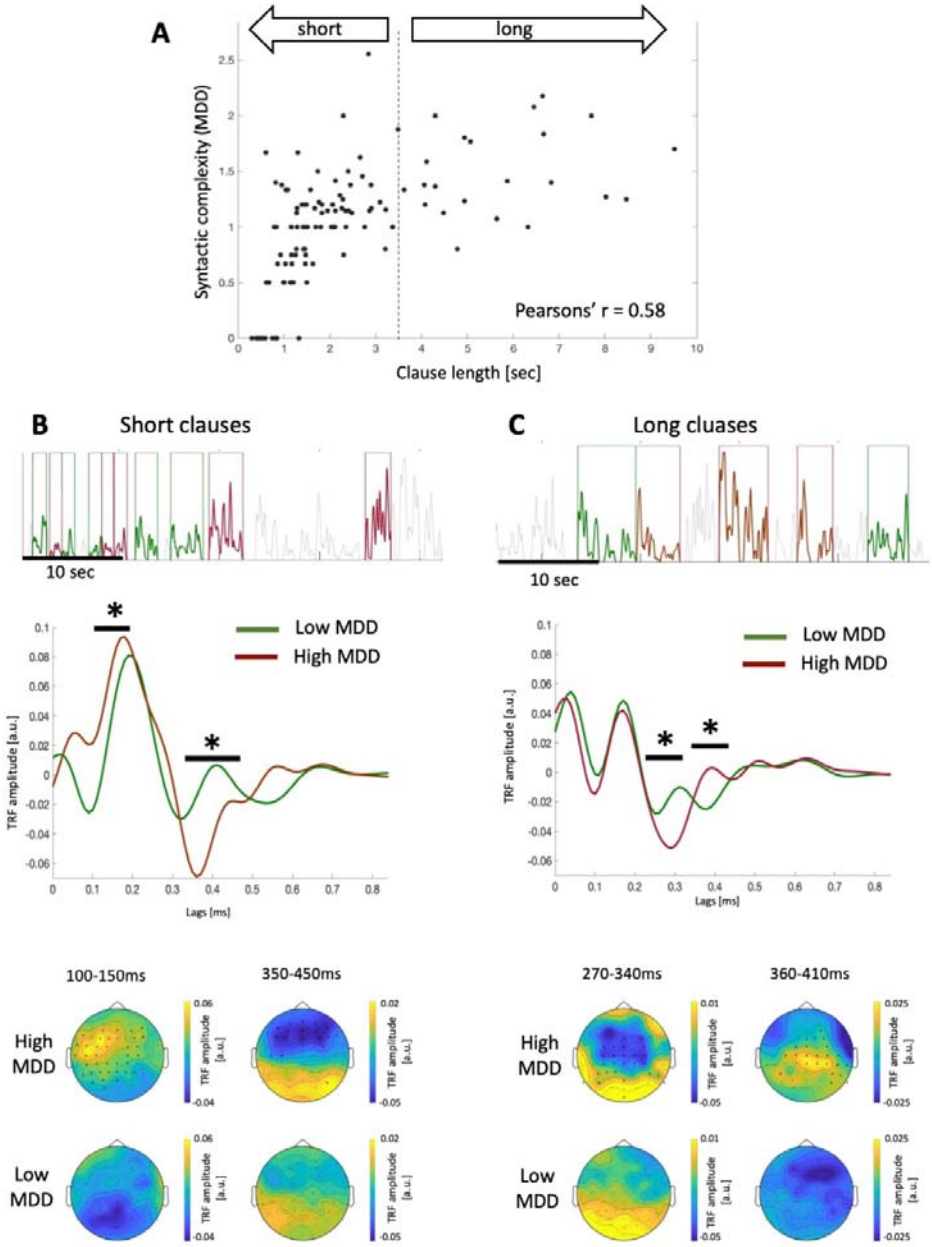

Figure 5. A. Scatter plot of clause-length and syntactic complexity (MDD) for all clauses in the stimulus, showing that the two are highly correlated. To avoid length-related confounds in the complexity analysis separated analyses of the effect of syntactic complexity were performed for short and long clauses (vertical line indicates group split). **B&C.** TRF analysis as a function of MDD, for short and long clauses separately. *Top:* A portion of the acoustic envelope (gray), overlayed with indication of clauses with low (green) and high (red) MDD. *Middle*: The TRF derived for clauses with low (green) and high MDD (red). Times with significant

differences between the TRFs are indicated with vertical lines and an asterisk ($p<0.05$ TCFE cluster corrected). *Bottom*: Scalp topographic of the TRF response in time-windows where significantly larger responses were found between clauses with low vs. high MDD. Electrodes where this effect was significant are indicated by asterisks on topography where the TRF response was highest.

Discussion

In the current study, we looked at how the neural speech tracking response is affected by different features of spontaneous speech. We focused specifically on (1) the presence of fillers; (2) detection of syntactic boundaries in continuous spontaneous speech and (3) the effect of syntactic complexity of clauses. The speech-tracking response, particularly the late negative peaking between 300-400 ms, was affected by all three features, supporting the utility of this metric for studying linguistic and other cognitive aspects of continuous speech processing.

Specifically, we found that fillers (e.g., "um", "uh", "like") elicited a weaker speech-tracking response than words that were part of the core/semantic meaning of the utterance. We also found differences in the TRF peaks for the first vs. last word in a phrase, providing a neural marker for on-line syntactic segmentation. Furthermore, the TRFs were modulated by syntactic complexity, with larger TRF peaks observed for more complex phrases. These findings offer important insights into the neural encoding of spontaneous speech, with all its complexities and imperfections. Before discussing each effect separately, and what they can teach us about real-life speech processing, it is worth considering how the estimated TRF speech tracking response relates to more traditional ERP measures capturing neural responses to language material.

ERP vs. TRF components – do they reflect the same processes?

The ERP world has a long tradition of associating different deflections in the signal ("components") to specific perceptual and/or cognitive operations, assigning meaning to both the amplitude and latency of these time-locked responses (Hagoort and Brown, 2000). The recent introduction of speech tracking methods and estimation of TRFs is often thought of as providing an equivalent to ERPs for continuous stimuli (Brodbeck and Simon, 2020; Dikker et al., 2020). Supporting this notion, TRFs to speech share many commonalities with auditory ERPs, which display similar time-courses of peak and troughs, and similar scalp-topographies and source estimations (Lalor et al., 2009). Early TRF components between 50-150 ms have been shown to be localized to auditory cortex and are primarily driven by the acoustics of the speech (Ding et al., 2014; Di Liberto et al., 2015). Importantly, these similarities between ERP and TRF components are not limited to sensory responses, but seem to extend to later components, which have been linked to higher-order cognitive and linguistic processing (Gwilliams and Davis, 2022). For example, linguistic and semantic features such of as predictability, surprise, lexical frequency and semantic, have been shown to affect TRF responses in time windows between 200-400ms (Brodbeck et al., 2018, 2022; Broderick et al., 2018, 2019; Donhauser and Baillet, 2020; Gillis et al., 2021), and have been localized to more frontal regions, in ways reminiscent of the P2 and N400 ERP responses to written words (Kutas and Federmeier, 2011). Hence, it is appealing to interpret the observed TRP responses as an extension of the vast ERP literature to the domain of continuous speech.

At the same time, we must be careful not to over-interpret these similarities for several reasons. First, the majority of ERP language studies quantify response to written material, rather than speech. Moreover, ERP studies use highly structured and controlled language materials, designed to test specific research questions, whereas the speech-materials used here and in the above-mentioned speech-tracking studies were not designed to probe specific semantic/linguistic questions, but these were derived post-hoc from a given stimulus. Hence,

the operationalization of ‘syntactic complexity’, ‘semantic predictability’, ‘lexical frequency’ etc. are different and less well-controlled when using pre-existing or unscripted continuous speech material. Another point to consider is that the peak latency of late components can vary substantially across studies – both for TRFs and ERPs. Along these lines, here we find a negative TRF peak peaking between 300-350 ms which is similar to the N400 ERP component, yet earlier than what is often reported. Whether these latency differences indicate that they capture different cognitive/linguistic processes or are due to incidental differences across studies (for example, due to parameter-choices) remains to be fully determined. Bearing these caveats in mind, and with the goal of gaining a unified perspective on how findings from more controlled experiments generalize to continuous real-life speech, we now turn to discuss the neural responses to features of spontaneous speech observed here.

Neural tracking of fillers

The comparison between the TRFs estimated for lexical words vs. fillers indicates a sharp difference in their encoding in the brain. Although fillers clearly serve a communicative function, and some have even evolved from words with lexical meaning (e.g., pragmatic markers that express reduced commitment, such as ‘like’, ‘you know’ (Ziv, 1988)), they do not contribute to the direct comprehension of the narrative and construction of semantic meaning. The substantially less prominent TRF response observed here for fillers vs. lexical words, starting with the earliest response at 50 ms, suggests that they are quickly identified as less important, and their sensory encoding is suppressed. This may be akin to modulations observed for task-irrelevant speech, for which reduced early TRF responses have been repeatedly shown (Ding and Simon, 2012; Zion Golumbic et al., 2013b; Holtze et al., 2021). Not surprisingly, the most prominent difference between responses to lexical words and fillers, even after controlling for their relative infrequency, was found for the negative response between 300-400 ms. As discussed above, this late TRF peak shares many

similarities with the ERP N400 response, which is linked to lexical access and semantic processing. These processes are less relevant for fillers, which is in line with the diminished response observed here.

This finding contributes to an ongoing debate about whether the insertion of fillers into spontaneous speech is a conscious choice of the speaker, just as it is for meaningful words, or if their production is more similar to the insertion of pauses, which are often simply a by-product of production delays (Clark and Fox Tree, 2002; Tian et al., 2017; Silber-Varod et al., 2021). To our knowledge, this is the first study to look directly at the neural response to fillers. Our results suggest that, at least from the listener's perspective, fillers are treated by the brain more like pauses than words. This pattern suggests that the brain applies online filtering to incoming speech, so as to "clean-up" the input from irrelevant insertions before applying syntactic and semantic integrative analysis at the sentence-level. It is still an open question whether this filtering mechanism is sensitive to the lexicality of the input or rather to the extent to which it contributes to meaning. For example, content words (nouns, verbs, adjectives) convey more information than function words (determiners, prepositions). Do we expect a similar effect? We leave this question for future research.

Syntactic boundaries in spontaneous speech

Listeners automatically segment streams of continuous speech into syntactic phrases that serve as basic processing units (Fodor and Bever, 1965; Garrett et al., 1966; Jarvella, 1971). Segmentation of continuous speech relies on two main features: the underlying syntactic structure of the utterance (Ding et al., 2016; Kaufeld et al., 2020; Har-shai Yahav and Zion Golumbic, 2021) and prosodic cues (Hirotani et al., 2006; Warren et al., 2009) (Langus et al., 2012; Hawthorne and Gerken, 2014). Since the current study uses spontaneous speech, these two features are inherently conflated, as prosodic breaks often align with

syntactic boundaries (Klatt, 1975; Cooper and Paccia-Cooper, 1980; Strangert, 1992; Strangert and Strangert, 1993). Consequently, we do not attempt to dissociate the independent contributions of prosodic processing and syntactic processing to online segmentation. Rather, our goal was to compare the neural response to opening vs. closing words of syntactic clauses, as a way to gauge the neural markers of detecting these boundaries 'on the fly', knowing that they are often accompanied by inflections of intonation and other prosodic cues (which were sometimes used to disambiguate the structure, see Methods). Two main differences were observed: closing words elicited larger positive response around 200 ms, reminiscent of the ERP P2 response, which was not observed for opening words. Additionally, the negative response between 300-400 ms was larger and earlier for opening vs. closing words. Below we discuss these results, on the background of parallel/similarly motivation ERP literature for individual words.

The increased P2-like response for closing words is in line with two effects that are sometimes found in reading studies for words at syntactic and/or prosodic boundaries. Sentence-final words have been shown to trigger a positive ERP response, although this has been seen in variable and somewhat later time windows ranging from 300 to 500 ms (Friedman et al., 1975; Kutas and Hillyard, 1980; Stowe et al., 2018). This ERP effect was seldom the primary focus of interest, and in fact researchers were often advised to avoid analyzing responses to final words, and hence, there is no direct investigation of this ERP positivity and its potential role at syntactic boundaries (see Stowe et al. 2018 for a review). That said, one interpretation of this positivity could be attributed to effects of syntactic integration, linguistic processes that do occur towards the end of a sentence and were associated with increased brain activation (Pallier et al. (2011), Nelson et al. (Nelson et al., 2017)). Accordingly, it is possible that the P2-like response observed here is a speech-processing correlate of the end-of-sentence responses known from these reading studies.

Another potentially related ERP effect that is not necessarily linguistically oriented, is the closure positive shift (CPS), a centro-parietal neural response observed at prosodic breaks (Steinhauer et al., 1999; Pannekamp et al., 2005; Peter et al., 2014). The latency of the CPS varies substantially across studies, and likely depends on the specific acoustic properties of the stimulus, yet it is consistently observed at clause boundaries where a comma would be expected (Steinhauer and Friederici, 2001). Importantly, although the CPS is a response to prosodic cues, it clearly interacts with syntactic boundaries (Kerkhofs et al., 2007), and can even be triggered by the mere expectation of a syntactic boundary even without a prosodic cue (Itzhak et al., 2010). Since the definition used here for 'closing' words corresponds to syntactic boundaries, that highly overlap with prosodic breaks, another possibility is that the observed P2-like response may be related to this CPS response. These two interpretations for the P2-like response are not mutually exclusive, and it is it likely that linguistic integrative processes that occur at syntactic boundaries interact highly with prosodic analysis of the input. There is clearly much more to be investigated regarding the on-line segmentation and syntactic analysis of natural speech, and it is hard to interpret our results given the exploratory nature of our study. However, beyond the current results, our study provides novel methods and promising directions to pursue these questions in spontaneous speech.

Besides the early positive effect, we also found that a later negative peak between 300-400 ms was larger and earlier for opening vs. closing words. We are not aware of previous studies that have directly made this comparison. However, bearing in mind the possible association between this response with the ERP N400 component, possible explanations for this effect may be related to the reduced semantic predictability of the first word or the reduced ability to apply integrative semantic processing at early stages of sentence processing. In line with this rationale, it has been shown that words that are less predictable evoke a larger N400 (Lau et al., 2013; Nieuwland et al., 2020). More particularly,

N400 response of opening words is reduced when they are more easily integrated with previous context (Calloway and Perfetti, 2017). However, additional work is required to better link the effects observed here for speech and the results of past reading studies.

Syntactic complexity in spontaneous speech

Sentences with more complex syntax incur a higher burden on cognitive processing. However, it is unclear how to investigate syntactic complexity's neural correlates with spontaneous speech. Past studies of syntactic complexity have mostly used highly controlled paradigms, comparing neural responses to sentences that differ from each other on only one linguistic dimension (Ferreira, 1991; Friederici et al., 2002; Ben-Shachar et al., 2003; Wingfield et al., 2003; Grodzinsky and Santi, 2008; Kaan et al., 2010). Within this controlled approach ('minimal linguistic pairs approach'), in the EEG domain, an ample of studies have successfully demonstrated increased N400 and P600 ERP components for higher-complexity sentences, mostly focusing on relative clauses and similar syntactic constructions such as WH-questions (Kluender and Kutas, 1993; King and Kutas, 1995; Müller et al., 1997; Fiebach et al., 2001, 2002; Friederici et al., 2002; Phillips et al., 2005; Ueno and Garnsey, 2008; Gouvea et al., 2010; Kaan et al., 2010; Kwon et al., 2013; Bulut et al., 2018).

Several metrics have been proposed in the past for quantifying syntactic complexity on a continuous basis (Gibson, 1998; Szmrecsanyi, 2004; Roark et al., 2007, 2011; Baumann, 2014; Jing and Liu, 2015; Ambati et al., 2016). In attempt to move towards studying language under more naturalistic conditions, such metrics can be useful when studying uncontrolled stimuli such as speech, as recently been shown by studies that associated such scores with fluctuations in the neural response (Brennan et al., 2016; Nelson et al., 2017; Lopopolo et al., 2021; Shain et al., 2021). Here we used the mean-dependency distance (MDD), computed per clause, which has been hypothesized to reflect working memory

demands (Liu, 2008; Gildea and Temperley, 2010; Collins, 2014; Futrell et al., 2015; Liu et al., 2017), yet to our knowledge, the neural correlates of MDD have not been tested before.

Syntactic analysis of spontaneous speech is further complicated by the presence of disfluencies and imperfections. However, despite the uncontrolled stimulus used here and the novel approach used for assessing the complexity of speech, we show that the speech tracking response to clauses was significantly modulated by their MDD, in a manner similar to the ERP pattern observed for highly controlled minimal pairs. The most striking effect was the enhanced negative response between 300-400 ms for clauses with high vs. low complexity, an effect that was consistent for both the short and long clauses. Syntactic complexity also showed additional effects, which varied for clauses of different lengths – with more complex clauses eliciting an enhanced P2-like response for short phrases and an enhanced later positive response for long phrases. However, further investigation is required into the potential interaction between complexity and clause length before these effects can be adequately interpreted.

As mentioned, MDD, like other distance metrics, may reflect the working memory cost required for keeping track of the syntactic structure as the sentence unfolds. Accordingly, the enhanced TRF responses to more complex clauses may reflect the higher burden on working memory that their analysis requires. In the ERP literature, N400 is rarely discussed in the context of working memory load (but see Salisbury, 2004), which suggests this TRF component may not be completely analogous to the N400. In addition, it is important to bear in mind that syntactic complexity is a multi-faceted phenomenon, with other components contributing to increased effort other than memory. For example, effort is required to merge words and phrases into larger hierarchical syntactic structures (Gibson, 1998; Brennan et al., 2016; Nelson et al., 2017). These aspects of syntactic complexity are not necessarily reflected by MDD, and could be distributed across different brain regions in

the language network (Lopopolo et al., 2021). It is beyond the scope of the current work to investigate the multiple facets of syntactic complexity. Nonetheless, the current findings provide compelling and novel evidence that phrase-level measures of complexity, such as long-distance dependencies, capture something essential about the cognitive and linguistic effort required for online speech processing.

Summary

As speech-processing research moves towards understanding how the brain encodes speech under increasingly realistic conditions, it is important to recognize that the type of speech we hear daily is linguistically imperfect. The current study offers an initial glimpse into some of the ways that these imperfections are dealt with in the neural response and demonstrates the robustness of the speech processing system for understanding spontaneous speech despite its disfluent and highly complex nature. Our results also point to a specific component of the speech-tracking response – a negative response akin to the N400 ERP response – that was modulated by all three factors studied here. The mechanistic and linguistic insights from this proof-of-concept study provide a basis for forming specific hypotheses for the continued investigation of how the brain deals with the natural imperfections of spontaneous speech.

Data availability statement

Data will be made available on OSF upon acceptance.

Acknowledgement

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projectnumber 490839860 (to MGB and EZG) and 411333557 (to MGB), and by the Israel Ministry of Science grant #88962 to EZG. We would like to thank Paz Har-shai Yahav, Rotem Budagzad-Jacobson, Orel Levi and Shiraz Levi for their assistance in speech annotation and data collection.

References

Ambati BR, Reddy S, Steedman M (2016) Assessing relative sentence complexity using an incremental ccg parser. In: Proceedings of NAACL-HLT, pp 1051–1057.

Arnold JE, Tanenhaus MK, Altmann RJ, Fagnano M (2004) The old and thee, uh, new: disfluency and reference resolution. Psychol Sci 15:578–582.

Bailey KGD, Ferreira F (2003) Disfluencies affect the parsing of garden-path sentences. J Mem Lang 49:183–200.

Barr DJ, Seyfeddinipur M (2010) The role of fillers in listener attributions for speaker disfluency. Lang Cogn Process 25:441–455.

Baumann P (2014) Dependencies and hierarchical structure in sentence processing. In: Proceedings of the Annual Meeting of the Cognitive Science Society, pp 36.

Ben-Shachar M, Hendler T, Kahn I, Ben-Bashat D, Grodzinsky Y (2003) The neural reality of syntactic transformations. Psychol Sci 14:433–440.

Blaauw E (1994) The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. Speech Commun 14:359–375.

Boersma P, Weenink D (2021) Praat: doing phonetics by computer.

Bortfeld H, Leon SD, Bloom JE, Schober MF, Brennan SE (2001) Disfluency rates in conversation: effects of age, relationship, topic, role, and gender. Lang Speech 44:123–147.

Brennan JR, Stabler EP, Van Wagenen SE, Luh W-M, Hale JT (2016) Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. Brain Lang 157:81–94.

Brennan SE, Schober MF (2001) How listeners compensate for disfluencies in spontaneous speech. J Mem Lang 44:274–296.

Brennan SE, Williams M (1995) The feeling of another's knowing: prosody and filled pauses

as cues to listeners about the metacognitive states of speakers. J Mem Lang 34:383–398.

Brodbeck C, Bhattasali S, Cruz Heredia AAL, Resnik P, Simon JZ, Lau E (2022) Parallel processing in speech perception with local and global representations of linguistic context. Elife 11.

Brodbeck C, Presacco A, Simon JZ (2018) Neural source dynamics of brain responses to continuous stimuli: speech processing from acoustics to comprehension. Neuroimage 172:162–174.

Brodbeck C, Simon JZ (2020) Continuous speech processing. Curr Opin Physiol 18:25–31.

Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr Biol 28:803-809.e3.

Broderick MP, Anderson AJ, Lalor EC (2019) Semantic context enhances the early auditory encoding of natural speech. J Neurosci 39:7564–7575.

Bulut T, Cheng SK, Xu KY, Hung DL, Wu DH (2018) Is there a processing preference for object relative clauses in chinese? evidence from erps. Front Psychol 9:995.

Calloway RC, Perfetti CA (2017) Integrative and predictive processes in text reading: the n400 across a sentence boundary. Lang Cogn Neurosci 32:1001–1016.

Caplan D, Alpert N, Waters G (1998) Effects of syntactic structure and propositional number on patterns of regional cerebral blood flow. J Cogn Neurosci 10:541–552.

Clark HH, Fox Tree JE (2002) Using uh and um in spontaneous speaking. Cognition 84:73–111.

Clark HH, Wasow T (1998) Repeating words in spontaneous speech. Cogn Psychol 37:201–242.

Collard P, Corley M, MacGregor LJ, Donaldson DI (2008) Attention orienting effects of hesitations in speech: evidence from erps. J Exp Psychol Learn Mem Cogn 34:696–702.

Collins MX (2014) Information density and dependency length as complementary cognitive models. J Psycholinguist Res 43:651–681.

Cooper WE, Paccia-Cooper J (1980) Syntax and Speech, 2013th ed. Cambridge, MA and London, England: Harvard University Press.

Corley M, MacGregor LJ, Donaldson DI (2007) It's the way that you, er, say it: hesitations in speech affect language comprehension. Cognition 105:658–668.

Corley M, Stewart OW (2008) Hesitation disfluencies in spontaneous speech: the meaning of um. Lang Linguist Compass 2:589–602.

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465.

Dikker S, Assaneo MF, Gwilliams L, Wang L, Kösem A (2020) Magnetoencephalography and language. Neuroimaging Clin 30:229–238.

Ding N, Chatterjee M, Simon JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage 88:41–46.

Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 19:158–164.

Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci 109:11854–11859.

Donhauser PW, Baillet S (2020) Two distinct neural timescales for predictive speech processing. Neuron 105:385-393.e9.

Ferreira F (1991) Effects of length and syntactic complexity on initiation times for prepared utterances. J Mem Lang 30:210–233.

Ferreira F, Patson ND (2007) The 'good enough' approach to language comprehension. Lang Linguist Compass 1:71–83.

Fiebach CJ, Schlesewsky M, Friederici AD (2001) Syntactic working memory and the

establishment of filler-gap dependencies: insights from erps and fmri. J Psycholinguist Res 2001 303 30:321–338.

Fiebach CJ, Schlesewsky M, Friederici AD (2002) Separating syntactic memory costs and syntactic integration costs during parsing: the processing of german wh-questions. J Mem Lang 47:250–272.

Fodor JA, Bever TG (1965) The psychological reality of linguistic segments. J Verbal Learning Verbal Behav 4:414–420.

Fox Tree JE (1995) The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. J Mem Lang 34:709–738.

Fox Tree JE (2001) Listeners' uses of um and uh in speech comprehension. Mem Cogn 29:320–326.

Fox Tree JE, Schrock JC (2002) Basic meanings of "you know" and "i mean." J Pragmat 34:727–747.

Fraundorf SH, Watson DG (2011) The disfluent discourse: effects of filled pauses on recall. J Mem Lang 65:161–175.

Friederici AD, Hahne A, Saddy D (2002) Distinct neurophysiological patterns reflecting aspects of syntactic complexity and syntactic repair. J Psycholinguist Res 31:45–63.

Friedman D, Simson R, Ritter W, Rapin I (1975) The late positive component (p300) and information processing in sentences. Electroencephalogr Clin Neurophysiol 38:255–262.

Futrell R, Mahowald K, Gibson E (2015) Large-scale evidence of dependency length minimization in 37 languages. Proc Natl Acad Sci U S A 112:10336–10341.

Garrett M, Bever T, Fodor J (1966) The active use of grammar in speech perception. Percept Psychophys 1966 11 1:30–32.

Gibson E (1998) Linguistic complexity: locality of syntactic dependencies. Cognition 68:1–76.

Gildea D, Temperley D (2010) Do grammars minimize dependency length? Cogn Sci 34:286–310.

Gillis M, Vanthornhout J, Simon JZ, Francart T, Brodbeck C (2021) Neural markers of speech comprehension: measuring eeg tracking of linguistic speech representations, controlling the speech acoustics. J Neurosci 41:10316–10329.

Gouvea AC, Phillips C, Kazanina N, Poeppel D (2010) The linguistic processes underlying the p600. Lang Cogn Process 25:149–188.

Grodzinsky Y, Santi A (2008) The battle for broca's region. Trends Cogn Sci 12:474–480.

Gwilliams L, Davis MH (2022) Extracting language content from speech sounds: the information theoretic approach. In: Speech Perception (Holt, L. L., Peelle JE, Coffin AB, Popper AN, Fay RR, eds), pp 113–139. Springer, Cham.

Hagoort P, Brown CM (2000) ERP effects of listening to speech: semantic erp effects. Neuropsychologia 38:1518–1530.

Har-shai Yahav P, Zion Golumbic E (2021) Linguistic processing of task-irrelevant speech at a cocktail party. Elife 10.

Haselow A (2017) Spontaneous spoken English: An integrated approach to the emergent grammar of speech (Kytö M, ed). Cambridge: Cambridge University Press.

Hawthorne K, Gerken LA (2014) From pauses to clauses: prosody facilitates learning of syntactic constituency. Cognition 133:420–428.

Hirotani M, Frazier L, Rayner K (2006) Punctuation and intonation effects on clause and sentence wrap-up: evidence from eye movements. J Mem Lang 54:425–443.

Holtze B, Jaeger M, Debener S, Adiloğlu K, Mirkovic B (2021) Are they calling my name? attention capture is reflected in the neural tracking of attended and ignored speech. Front Neurosci 15:234.

Hudson R (1995) Measuring syntatic difficulty. Manuscript.

Inbar M, Grossman E, Landau AN (2020) Sequences of intonation units form a ~ 1 hz rhythm. Sci Rep 10:1–9.

Itzhak I, Pauker E, Drury JE, Baum SR, Steinhauer K (2010) Event-related potentials show online influence of lexical biases on prosodic processing. Neuroreport 21:8–13.

Jarvella RJ (1971) Syntactic processing of connected speech. J Verbal Learning Verbal Behav 10:409–416.

Jing Y, Liu H (2015) Mean hierarchical distance augmenting mean dependency distance. In: Proceedings of the third international conference on dependency linguistics (Depling 2015), pp 161–170.

Just MA, Carpenter PA (1980) A theory of reading: from eye fixations to comprehension. Psychol Rev 87:329–354.

Kaan E, Harris A, Gibson E, Holcomb P (2010) The p600 as an index of syntactic integration difficulty. Lang Cogn Process 15:159–201.

Kaufeld G, Bosker HR, ten Oever S, Alday PM, Meyer AS, Martin AE (2020) Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. J Neurosci:JN-RM-0302-20.

Keitel A, Gross J, Kayser C (2018) Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLOS Biol 16:e2004473.

Kerkhofs R, Vonk W, Schriefers H, Chwilla DJ (2007) Discourse, syntax, and prosody: the brain reveals an immediate interaction. J Cogn Neurosci 19:1421–1434.

King JW, Kutas M (1995) Who did what and when? using word- and clause-level erps to monitor working memory usage in reading. J Cogn Neurosci 7:376–395.

Klatt DH (1975) Vowel lengthening is syntactically determined in a connected discourse. J Phon 3:129–140.

Kluender R, Kutas M (1993) Bridging the gap: evidence from erps on the processing of

unbounded dependencies. J Cogn Neurosci 5:196–214.

Kutas M, Federmeier KD (2011) Thirty years and counting: finding meaning in the n400 component of the event related brain potential (erp). Annu Rev Psychol 62:621–647.

Kutas M, Hillyard SA (1980) Reading senseless sentences: brain potentials reflect semantic incongruity. Science (80-) 207:203–205.

Kwon N, Kutas M, Kluender R, Polinsky M (2013) Subject/object processing asymmetries in korean relative clauses: evidence from erp data. Language (Baltim) 89:537.

Lalor EC, Power AJ, Reilly RB, Foxe JJ (2009) Resolving precise temporal processing properties of the auditory system using continuous stimuli. J Neurophysiol 102:349–359.

Langus A, Marchetto E, Bion RAH, Nespor M (2012) Can prosody be used to discover hierarchical structure in continuous speech? J Mem Lang 66:285–306.

Lau E, Tanaka N (2021) The subject advantage in relative clauses: a review. Glossa 6.

Lau EF, Holcomb PJ, Kuperberg GR (2013) Dissociating n400 effects of prediction from association in single-word contexts. J Cogn Neurosci 25:484–502.

Lau EF, Phillips C, Poeppel D (2008) A cortical network for semantics: (de)constructing the n400. Nat Rev Neurosci 9:920–933.

Liberman MC (1982) The cochlear frequency map for the cat: labeling auditory nerve fibers of known characteristic frequency. J Acoust Soc Am 72:1441–1449.

Liu H (2008) Dependency distance as a metric of language comprehension difficulty. J Cogn Sci (Seoul) 9:159–191.

Liu H, Xu C, Liang J (2017) Dependency distance: a new perspective on syntactic patterns in natural languages. Phys Life Rev 21:171–193.

Liu Y, Chawla N V., Harper MP, Shriberg E, Stolcke A (2006) A study in machine learning from imbalanced data for sentence boundary detection in speech. Comput Speech Lang

20:468–494.

Lopopolo A, van den Bosch A, Petersson K-M, Willems RM (2021) Distinguishing syntactic operations in the brain: dependency and phrase-structure parsing. Neurobiol Lang 2:152–175.

Mehta G, Cutler A (1988) Detection of target phonemes in spontaneous and read speech. Lang Speech 31:135–156.

More A, Seker A, Basmova V, Tsarfaty R (2019) Joint transition-based models for morpho-syntactic parsing: parsing strategies for mrls and a case study from modern hebrew. Trans Assoc Comput Linguist 7:33–48.

Müller HM, King JW, Kutas M (1997) Event-related potentials elicited by spoken relative clauses. Cogn Brain Res 5:193–203.

Nelson MJ, El Karoui I, Giber K, Yang X, Cohen L, Koopman H, Cash SS, Naccache L, Hale JT, Pallier C, Dehaene S (2017) Neurophysiological dynamics of phrase-structure building during sentence processing. Proc Natl Acad Sci U S A 114:E3669–E3678.

Nieuwland MS et al. (2020) Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. Philos Trans R Soc B 375.

Obleser J, Kayser C (2019) Neural entrainment and attentional selection in the listening brain. Trends Cogn Sci 23:913–926.

Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. Comput Intell Neurosci 2011.

Pallier C, Devauchelle A-D, Dehaene S (2011) Cortical representation of the constituent structure of sentences. Proc Natl Acad Sci U S A 108:2522–2527.

Pannekamp A, Toepel U, Alter K, Hahne A, Friederici AD (2005) Prosody-driven sentence

processing: an event-related brain potential study. J Cogn Neurosci 17:407–421.

Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK (2019) PsychoPy2: experiments in behavior made easy. Behav Res Methods 2019 511 51:195–203.

Peter V, McArthur G, Crain S (2014) Using event-related potentials to measure phrase boundary perception in english. BMC Neurosci 15:1–11.

Phillips C, Kazanina N, Abada SH (2005) ERP effects of the processing of syntactic long-distance dependencies. Cogn Brain Res 22:407–428.

Roark B, Mitchell M, Hollingshead K (2007) Syntactic complexity measures for detecting mild cognitive impairment. In: BioNLP 2007: Biological, translational, and clinical language processing, pp 1–8.

Roark B, Mitchell M, Hosom J-P, Hollingshead K, Kaye J (2011) Spoken language derived measures for detecting mild cognitive impairment. In: IEEE Transactions on Audio, Speech and Language Processing, pp 2081–2090. IEEE.

Salisbury DF (2004) Semantic memory and verbal working memory correlates of n400 to subordinate homographs. Brain Cogn 55:396–399.

Shain C, Blank IA, Fedorenko E, Gibson E, Schuler W (2021) Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. bioRxiv:2021.09.18.460917.

Shriberg E (2001) To ‘errrr’ is human: ecology and acoustics of speech disfluencies. J Int Phon Assoc 31:153–169.

Silber-Varod V, Gósy M, Lerner A (2021) Is it a filler or a pause? a quantitative analysis of filled pauses in hebrew. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 12997 LNAI:638–648.

Smith VL, Clark HH (1993) On the course of answering questions. J Mem Lang 32:25–38.

Steinhauer K, Alter K, Friederici AD (1999) Brain potentials indicate immediate use of prosodic cues in natural speech processing. Nat Neurosci 2:191–196.

Steinhauer K, Friederici AD (2001) Prosodic boundaries, comma rules, and brain responses: the closure positive shift in erps as a universal marker for prosodic phrasing in listeners and readers. J Psycholinguist Res 30:267–295.

Stowe LA, Kaan E, Sabourin L, Taylor RC (2018) The sentence wrap-up dogma. Cognition 176:232–247.

Strangert E (1992) Prosodic cues to the perception of syntactic boundaries. In: 2nd International Conference on Spoken Language Processing (ICSLP 92), pp 1283–1285.

Strangert E, Strangert B (1993) Prosody in the perception of syntactic boundaries. In: EUROSPEECH ’93: Third European Conference on Speech Communication and Technology, pp 1209–1210.

Szmrecsanyi B (2004) On operationalizing syntactic complextity. JADT 2004 7es Journées Int d’Analyse Stat des Données Textuelles:1031–1038.

Tian Y, Maruyama T, Ginzburg J (2017) Self addressed questions and filled pauses: a cross-linguistic investigation. J Psycholinguist Res 46:905–922.

Tottie G (2014) On the use of uh and um in american english. Funct Lang 21:6–29.

Traxler MJ (2014) Trends in syntactic parsing: anticipation, bayesian estimation, and good-enough parsing. Trends Cogn Sci 18:605–611.

Ueno M, Garnsey SM (2008) An erp study of the processing of subject and object relative clauses in japanese. Lang Cogn Process 23:646–688.

Warren T, White SJ, Reichle ED (2009) Investigating the causes of wrap-up effects: evidence from eye movements and e–z reader. Cognition 111:132–137.

Watanabe M, Hirose K, Den Y, Minematsu N (2008) Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. Speech Commun 50:81–94.

Wieling M, Grieve J, Bouma G, Fruehwald J, Coleman J, Liberman M (2016) Variation and change in the use of hesitation markers in germanic languages. Lang Dyn Chang 6:199–234.

Wingfield A, Peelle JE, Grossman M (2003) Speech rate and syntactic complexity as multiplicative factors in speech comprehension by young and older adults. Aging, Neuropsychol Cogn 10:310–322.

Zion Golumbic EM, Cogan GB, Schroeder CE, Poeppel D (2013a) Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party.” J Neurosci 33.

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013b) Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” Neuron 77:980–991.

Ziv Y (1988) Lexical hedges and non-committal terms. Acta Linguist Hung 38:261–274.