Behavioral/Cognitive

Linguistic Structure and Meaning Organize Neural Oscillations into a Content-Specific Hierarchy

Greta Kaufeld,[1] Hans Rutger Bosker,[1,2] Sanne ten Oever,[1,2] Phillip M. Alday,[1] Antje S. Meyer,[1,2] and Andrea E. Martin[1,2]

[1]Max Planck Institute for Psycholinguistics, 6525 XD, Nijmegen, The Netherlands, and [2]Donders Institute, Radboud University, 6525EN Nijmegen, The Netherlands

Neural oscillations track linguistic information during speech comprehension (Ding et al., 2016; Keitel et al., 2018), and are known to be modulated by acoustic landmarks and speech intelligibility (Doelling et al., 2014; Zoefel and VanRullen, 2015). However, studies investigating linguistic tracking have either relied on non-naturalistic isochronous stimuli or failed to fully control for prosody. Therefore, it is still unclear whether low-frequency activity tracks linguistic structure during natural speech, where linguistic structure does not follow such a palpable temporal pattern. Here, we measured electroencephalography (EEG) and manipulated the presence of semantic and syntactic information apart from the timescale of their occurrence, while carefully controlling for the acoustic-prosodic and lexical-semantic information in the signal. EEG was recorded while 29 adult native speakers (22 women, 7 men) listened to naturally spoken Dutch sentences, jabberwocky controls with morphemes and sentential prosody, word lists with lexical content but no phrase structure, and backward acoustically matched controls. Mutual information (MI) analysis revealed sensitivity to linguistic content: MI was highest for sentences at the phrasal (0.8–1.1 Hz) and lexical (1.9–2.8 Hz) timescales, suggesting that the delta-band is modulated by lexically driven combinatorial processing beyond prosody, and that linguistic content (i.e., structure and meaning) organizes neural oscillations beyond the timescale and rhythmicity of the stimulus. This pattern is consistent with neurophysiologically inspired models of language comprehension (Martin, 2016, 2020; Martin and Doumas, 2017) where oscillations encode endogenously generated linguistic content over and above exogenous or stimulus-driven timing and rhythm information.

Key words: combinatorial processing; lexical semantics; mutual information; neural oscillations; prosody; sentence comprehension

Significance Statement

Biological systems like the brain encode their environment not only by reacting in a series of stimulus-driven responses, but by combining stimulus-driven information with endogenous, internally generated, inferential knowledge and meaning. Understanding language from speech is the human benchmark for this. Much research focuses on the purely stimulus-driven response, but here, we focus on the goal of language behavior: conveying structure and meaning. To that end, we use naturalistic stimuli that contrast acoustic-prosodic and lexical-semantic information to show that, during spoken language comprehension, oscillatory modulations reflect computations related to inferring structure and meaning from the acoustic signal. Our experiment provides the first evidence to date that compositional structure and meaning organize the oscillatory response, above and beyond prosodic and lexical controls.

Introduction

How the brain maps the acoustics of speech onto abstract structure and meaning during spoken language comprehension remains a core question across cognitive science and neuroscience. A large body of research has shown that neural populations closely track the envelope of the speech signal, which correlates with the syllable rate (Peelle and Davis, 2012; Zoefel and VanRullen, 2015; Kösem et al., 2018), yet much less is known about the degree to which neural responses encode higher-level linguistic information such as words, phrases, and clauses. While previous studies suggest a crucial role for delta-band oscillations in the top-down generation of hierarchically

Received Feb. 11, 2020; revised Sep. 25, 2020; accepted Oct. 3, 2020.

Author contributions: G.K. and A.E.M. designed research; G.K. and A.E.M. performed research; G.K., H.R.B., S.t.O., P.M.A., and A.E.M. analyzed data; G.K., A.S.M., and A.E.M. wrote the paper.

The authors declare no competing financial interests.

A.E.M. was supported by the Max Planck Research Group "Language and Computation in Neural Systems" and by the Netherlands Organization for Scientific Research (Grant 016.Vidi.188.029). We thank Anne Keitel for sharing her expertise and scripts for the Mutual Information analysis; Laurel Brehm for statistical advice; Joe Rodd and Merel Maslowski for help with the experiment visualization (Fig. 1); Annelies van Wijngaarden for lending her voice for stimulus recording; Karthikeya Kaushik, Zina al-Jibouri, and Dylan Opdam for research assistance; and Micha Heilbron for feedback on an earlier draft of this manuscript.

Correspondence should be addressed to Andrea E. Martin at andrea.martin@mpi.nl.

https://doi.org/10.1523/JNEUROSCI.0302-20.2020

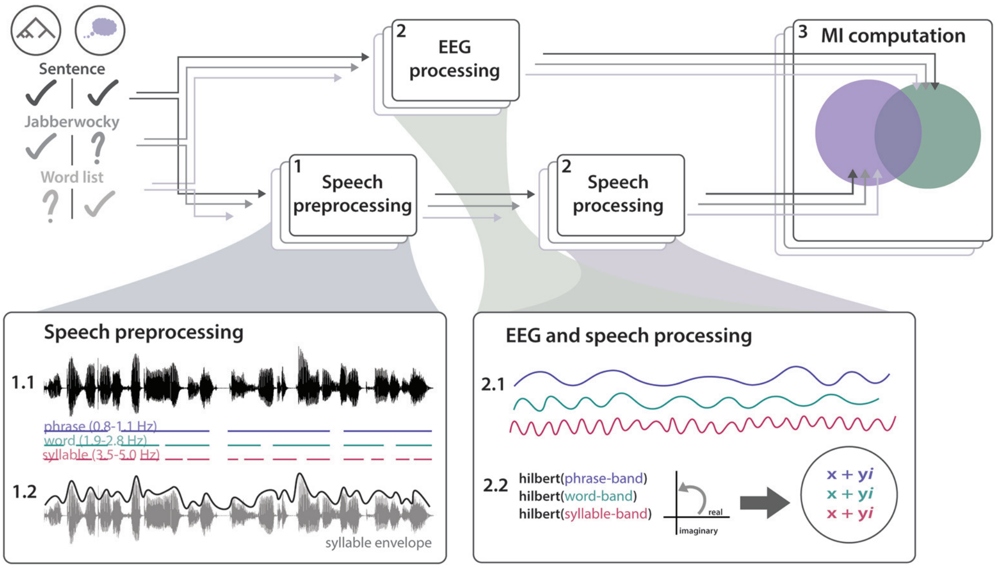

Figure 1. Experimental design and analysis pipeline. Participants listened to sentences, jabberwocky items, and word lists while their brain response was recorded using EEG. Step 1: Speech Processing: 1.1, the speech signal is annotated for the occurrence of phrases, words, and syllables in the stimuli, and, based on this, frequency bands of interest for each of the linguistic units can be identified; 1.2, a cochlear filter is applied to the speech stimuli and the amplitude envelope is extracted. Step 2: further processing is identical for both speech and EEG modalities: 2.1, broadband filters are applied in the previously identified frequency bands of interest; 2.2, Hilbert transforms are computed in each filtered signal, and real and imaginary parts of the Hilbert transform output are used for further analysis. Step 3: MI computation; mutual information is computed between the preprocessed speech and EEG signal in each of the three conditions and their respective backward controls.

structured linguistic representations (Ding et al., 2016; Keitel et al., 2018), they have so far either relied on non-naturalistic stimuli or failed to fully control for prosody. Here, we use a novel experimental design that allows us to investigate how structure and meaning shape the tracking of higher-level linguistic units, while using naturalistic stimuli and carefully controlling for prosodic fluctuations.

The strongest evidence for tracking of linguistic information so far are studies by Ding et al. (2016, 2017a), who found enhanced activity in the delta frequency range for sentences compared with word lists. They investigated this using isochronous, synthesized stimuli devoid of prosodic information. Yet phrases, clauses, and sentences usually do have acoustic-prosodic correlates (e.g., pauses, intonational contours, final lengthening, fundamental frequency reset; Eisner and McQueen, 2018). These might not be as prominent in the modulation spectrum of speech as syllables (Ding et al., 2017b), but listeners draw on them during language comprehension and learning (Soderstrom et al., 2003). As such, Ding et al. (2017a) cannot clearly distinguish between the generation of linguistic structure and meaning versus inferred prosody, and it is unclear whether their results generalize to naturalistic stimuli, where the timing of linguistic units is more variable.

Almost orthogonally to Ding et al. (2016, 2017a), Keitel et al. (2018) used naturalistic stimuli and found enhanced tracking (compared with reversed controls) in the delta-theta frequency range. However, as they did not include a systematic control for linguistic content, it is unclear whether their results are driven by tracking of prosodic information in the acoustic signal, rather than linguistic information.

In the current study, we bridge this gap by contrasting these two core sources of linguistic representations: prosodic structure, which can, but does not always, correlate with syntactic and information structure, and lexical semantics, which arises in isolated words and concepts. Participants listened to naturally spoken, structurally homogenous sentences, jabberwocky items (containing sentence-like prosody, but no lexical semantics), and word lists (containing lexical semantics, but no sentence-like structure and prosody; see Table 1 for examples). Additionally, we used reversed speech as the core control of our experiment because it has an identical modulation spectrum for each forward condition.

Using electroencephalography (EEG), we analyzed tracking at linguistically relevant timescales as quantified by mutual information (MI)—a typical measure of neural tracking that captures the informational similarity between two signals (Cogan and Poeppel, 2011; Gross et al., 2013; Kayser et al., 2015; Keitel et al., 2017, 2018). Figure 1 shows an overview of the experimental design and analysis pipeline.

We hypothesize that neural tracking ("entrainment in the broad sense," as defined by Obleser and Kayser, 2019) will be stronger for stimuli containing higher-level linguistic structure and meaning, above and beyond the acoustic-prosodic (jabberwocky) and lexical-semantic (word list) controls. This may reflect a process of perceptual inference (Martin, 2016, 2020), whereby biological systems like the brain encode their environment not only by reacting in a series of stimulus-driven responses, but by combining stimulus-driven information with endogenous, internally generated, inferential knowledge and meaning (Meyer et al., 2019). In sum, our study offers novel insights into how structure and meaning influence the neural response to natural speech above and beyond prosodic modulations and word-level meaning.

Materials and Methods

Participants. Thirty-five native Dutch speakers (26 females, 9 males; age range, 19–32 years; mean age, 23 years) participated in the experiment. They were recruited from the Max Planck Institute for Psycholinguistics (MPI) participant database with written consent approved by the Ethics Committee of the Social Sciences Department of Radboud University (Project code: ECSW2014-1003-196a). Six participants were excluded from the analysis because of excessive artifact

Table 1. Example items in Sentence, Jabberwocky, and Wordlist conditions

Sentence	Jabberwocky	Wordlist
[Bange helden] [plukken bloemen] en de [bruine vogels] [halen takken]	[Garge ralden] [spunken drijmen] en de [druize gomels] [paven mukken]	[helden bloemen] [vogels takken] de en [plukken halen] [bange bruine]
[*Timid heroes*] [*pluck flowers*] *and the* [*brown birds*] [*gather branches*]	[*Flimid lerops*] [*bruck clowters*] *and the* [*trown plirds*] [*shmather blamches*]	[*heroes flowers*] [*birds branches*] *the and* [*pluck gather*] [*timid brown*]

Sentences consisted of 10 words [disyllabic, except for "de" ("*the*") and "en" ("*and*")] and carried sentence prosody. Jabberwocky items consisted of 10 pseudowords with morphology; they also carried sentence prosody. Word lists consisted of the same 10 words as the sentence condition, but scrambled so as to be syntactically implausible. They had list prosody. Marked with square brackets are "phrases" in all three conditions. Note that the pseudowords/words in all three conditions had the same stress patterns.

contamination. All participants in the experiment reported normal hearing and were remunerated for their participation.

Materials. The experiment used the following three conditions: Sentence, Jabberwocky, and Wordlist. Eighty sets (triplets) of the three conditions (Sentence, Jabberwocky, Wordlist) were created, resulting in 240 stimuli. In addition to one "standard" forward presentation of each stimulus, participants also listened to a version of each of the stimuli played backward, thus resulting in a total of 480 stimuli.

Dutch stimuli consisted of 10 words, which were all disyllabic except for "de" (*the*) and "en" (*and*), thus resulting in 18 syllables in total. Sentences all consisted of two coordinate clauses, which followed the structure [*Adj N V N Conj Det Adj N V N*]. Word lists consisted of the same 10 words as in the Sentence condition, but scrambled in syntactically implausible ways (either [*V V Adj Adj Det Conj N N N N*], or [*N N N N Det Conj V V Adj Adj*], to avoid any plausible internal combinations of words). Jabberwocky items were created using the wuggy pseudoword generator (Keuleers and Brysbaert, 2010), following the same syntactic structure as the Sentences. Specifically, standard wuggy parameters were set to match two of three subsyllabic segments wherever possible, as well as letter length, transition frequencies, and length of subsyllabic segments. The lexicality feature of wuggy was used to ensure that none of the generated pseudowords were existing lexical items in Dutch. In addition, all pseudowords were proofread by native Dutch speakers to ensure that none of their phonetic forms matched that of an existing word in Dutch. Inflectional morphemes (e.g., plural morphemes) as well as function words ("de" - *the* and "en" - *and*) were kept unchanged. Table 1 shows an example of stimuli in each condition. (Please see https://osf.io/rv5y7/ for a list of all 480 stimuli and their translations.)

Forward stimuli were recorded by a female native speaker of Dutch in a sound-attenuating recording booth. All stimuli were recorded at a sampling rate of 44.1 kHz (mono), using the Audacity sound recording and analysis software (Audacity Team, 2014). After recording, pauses were normalized to $\sim$150 ms in all stimuli, and the intensity was scaled to 70 dB using the Praat voice analysis software (Boersma and Weenink, 2020). Stimuli from all three conditions were then reversed using Praat. Figure 2 shows modulation spectra for forward and backward conditions.

Procedure. Participants were tested individually in a sound-attenuating and Faraday cage-enclosed booth. They first completed a practice session with four trials (one from each forward condition and one backward example) to become familiarized with the experiment. All 80 stimuli from each condition were presented to the participants in separate blocks. The order of the blocks was pseudorandomized across listeners, and the order of the items within each block was randomized. During each trial, participants were instructed to look at a fixation cross, which was displayed at the center of the screen (to minimize eye movements during the trial), and listen to the audio, which was presented to them at a comfortable level of loudness. The audio recording was presented 500 ms after the fixation cross appeared on the screen, and the fixation cross remained on the screen for the entire duration of the audio recording. Fifty milliseconds after the end of each recording, the screen changed to a transition screen [a series of hash symbols (#####) indicating that participants could blink and briefly rest their eyes], after which participants could advance to the next item via a button-press. After each block, participants were allowed to take a self-paced break. The experiment was run using the Presentation software (Neurobehavioral Systems) and took $\sim$50–60 min to complete. EEG was continuously recorded with a 64-channel EEG system (MPI equidistant montage) connected to a BrainAmp amplifier using BrainVision Recorder software, digitized at a sampling rate of 500 Hz and referenced to the left mastoid. The time constant for the hardware high-pass filter was 10 s (0.016 Hz; first-order Butterworth filter with 6 dB/octave), the high-cutoff frequency was 249 Hz. The impedance of electrodes was kept at $<$25 kΩ. Data were rereferenced offline to the average reference.

EEG data preprocessing. The analysis steps were conducted using the FieldTrip Analysis Toolkit revision 20180320 (Oostenveld et al., 2011) on MATLAB version 2016a (MathWorks). The raw EEG signal was segmented into a series of variable length epochs, starting at 200 ms before the onset of the utterance and lasting until 200 ms after its end. The signal was low-pass filtered to 70 Hz, and a bandstop filter centered at $\sim$50 Hz ($\pm$2 Hz) was applied in each epoch to exclude line noise [both zero-phase FIR (finite impulse response) filters using Hamming windows]. All data were visually inspected, and channels contaminated with excessive noise were excluded from the analysis. Independent component analysis was performed on the remaining channels, and components related to eye movements, blinking, or motion artifacts were subtracted from the signal. Epochs containing voltage fluctuations exceeding $\pm$100 μV or exceeding a range of 150 μV were excluded from further analysis. We selected a cluster of 22 electrodes for all further analyses based on previous studies that found broadly distributed effects related to sentence processing (Kutas and Federmeier, 2000; Kutas et al., 2006; see also Ding et al., 2017a). Specifically, the electrode selection included the following electrodes: 1, 2, 3, 4, 5, 8, 9, 10, 11, 28, 29, 30, 31, 33, 34, 35, 36, 37, 40, 41, 42, and 43 (electrode names based on the MPI equidistant layout). We note that our results also hold for all electrodes, as described in the Results section below.

Speech preprocessing. For each stimulus, we computed the wideband speech envelope at a sampling rate of 150 Hz following the procedure reported by Keitel et al. (2018) and others (Gross et al., 2013; Keitel et al., 2017). We first filtered the acoustic waveforms into eight frequency bands (100–8000 Hz; third-order Butterworth filter, forward and reverse), equidistant on the cochlear frequency map (Smith et al., 2002). We then estimated the wideband speech envelope by computing the magnitude of the Hilbert transformed signal in each band and averaging across bands.

The timescales of interest for further mutual information analysis were identified in a fashion similar to that described in the study by Keitel et al. (2018). We first annotated the occurrence of linguistic units (phrases, words, and syllables) in the speech stimuli. Here, phrases were defined as adjective-noun/noun-verb combinations (e.g., in the Sentence condition: "bange helden" – *timid heroes*; "plukken bloemen" – *pluck flowers,* and so on; in the Jabberwocky condition: "garge ralden" – *flimid lerops* etc.; in the Wordlist condition, a "pseudo-phrase" corresponds to adjacent noun–noun, verb–verb, and adjective–adjective pairs; e.g., "helden bloemen" – *heroes flowers*). Unit-specific bands of interest were then identified by converting each of the rates into frequency ranges across conditions. This resulted in the following bands: 0.8–1.1 Hz (phrases); 1.9–2.8 Hz (words); and 3.5–5.0 Hz (syllables). Note that the problem the brain faces during spoken language comprehension is even more complex than this, because the timescales of linguistic units can highly overlap, even within a single sentence (Obleser et al., 2012). Populations of neurons that "entrain" to words will thus also have to be sensitive to information that occurs outside of these—rather narrow—frequency bands.

For an additional, exploratory annotation-based MI analysis (see Results, subsection Tracking of abstract linguistic units), we further created linguistically abstracted versions of our stimuli. Specifically, our aim was to create annotations that captured linguistic information at the phrase frequency entirely independent of the acoustic signal. Based on

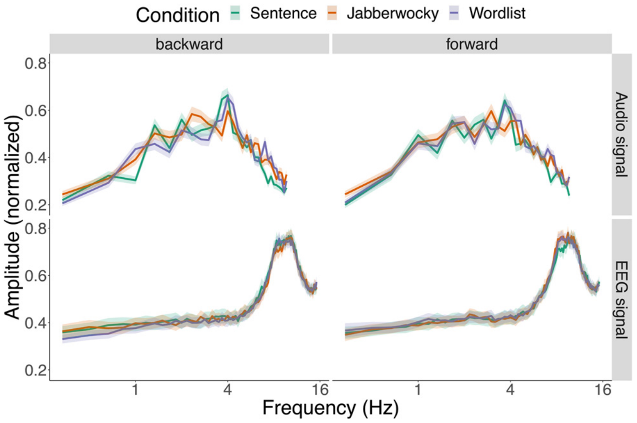

Figure 2. Modulation spectra of forward and backward stimuli. Green, Sentence; orange, Jabberwocky; purple, Wordlist. Modulation spectra were calculated following the procedure and MATLAB script described in the study by Ding et al. (2017b). Note that a cochlear filter is applied to the acoustic stimuli, but not the brain data. Small deviations between the modulation spectrum of each forward condition and its backward counterpart are because of numerical inaccuracy; mathematically, the frequency components of forward and backward stimuli are identical.

the word-level annotations of our stimuli, we created dimensionality-reduced arrays for further analysis (see the "Semantic composition" analyses reported by Brodbeck et al., 2018). Specifically, we identified all time points in the spoken materials where words could be integrated into phrases and marked each of these words associated with phrase composition [e.g., in a sentence such as "bange helden plukken bloemen en de bruine vogels halen takken" (*timid heroes pluck flowers and the brown birds gather branches*), the words "helden" (*heroes*), "bloemen" (*flowers*), "vogels" (*birds*), and "takken" (*branches*) were marked]. All these critical words were coded as 1 for their entire duration, while all other timepoints (samples) were marked as 0 (Brodbeck et al., 2018). This resulted in an abstract "spike train" array of phrase-level structure building that is independent of the acoustic envelope. We repeated this procedure for all items individually in all three conditions, since our stimuli were naturally spoken and thus differed slightly in duration and time course. Note that, consequently, this "phrase-level composition array" is somewhat arbitrary for the Wordlist condition as there are, per definition, no phrases in a word list. We annotated "pseudo-phrases" the same way as shown in Table 1. The procedure is visualized in Figure 3.

Mutual information analysis. We used MI to quantify the statistical dependency between the speech envelopes and the EEG recordings according to the procedure described in the study by Keitel et al. (2018; see also Gross et al., 2013; Kayser et al., 2015; Keitel et al., 2017). Based on the previously identified frequency bands of interest (see subsection "Speech preprocessing" above), we filtered both speech envelopes and EEG signals in each band (third-order Butterworth filter, forward and reverse). We then computed the Hilbert transform in each band, which resulted in two sets of two-dimensional variables (one for speech signals and one for EEG responses) in each condition (forward and backward; see Ince et al., 2017; for a more in-depth description). To take the brain–stimulus lag into account, we computed MI at five different lags, ranging from 60 to 140 ms in steps of 20 ms, and to exclude strong auditory-evoked responses to the onset of auditory stimulation in each trial, we excluded the first 200 ms of each stimulus–signal pair. MI values from all five lags were averaged for subsequent statistical evaluation. We further concatenated all trials from speech and brain signals to increase the robustness of MI computation (Keitel et al., 2018). In addition to computing "general" MI (containing information about both phase and power), we also isolated the part of the Hilbert transform corresponding to phase and computed "phase MI" values, separately.

Statistical analysis. To test whether the statistical dependency between the speech envelope and the EEG data as captured by MI was modulated by the linguistic structure and content of the stimulus, we compared MI values in all three frequency bands separately. Linear mixed models were fitted to the log-transformed, trimmed (5% on each end of the distribution) MI values in each frequency band using lme4 (Bates et al., 2015) in R (R Core Team, 2018). Models included main effects of Condition (three levels: Sentence, Wordlist, Jabberwocky) and Direction (two levels: Forward, Backward), as well as their interaction. All models included by-participant random intercepts and random slopes for the Condition * Direction interaction. For model coefficients, degrees of freedom were approximated using Satterthwaite's method, as implemented in the package lmerTest (Kuznetsova et al., 2017). We used treatment coding in all models, with Sentence being the reference level for Condition, and Forward the reference level for Direction. We then computed all pairwise comparisons within each direction using estimated marginal means (Tukey's correction for multiple comparisons) with emmeans (Length, 2018) in R (i.e., comparing Sentence Forward to Jabberwocky Forward and Wordlist Forward, but never Sentence Forward to Jabberwocky Backward, because we had no hypotheses about these comparisons). The same statistical analyses, including identical model structures, were further applied to MI values computed on the isolated phase coefficients.

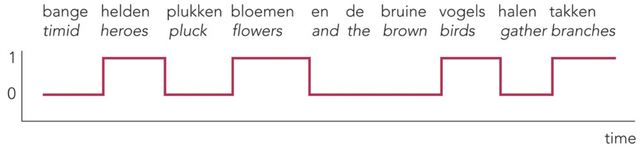

Figure 3. Visualization of the phrase-level annotations (inspired by Brodbeck et al., 2018, their Fig. 2). Across time, the response array takes value 0 for words that cannot (yet) be integrated into phrases, and value 1 for words that can, resulting in a "pulse train" array.

For the exploratory dimensionality-reduced MI analysis, we performed the same set of statistical analyses (but only in one single-frequency band). Specifically, we fitted a linear mixed model including main effects of Condition (three levels: Sentence, Wordlist, Jabberwocky) and Direction (two levels: Forward, Backward), as well as their interaction and by-participant random intercepts and random slopes for the Condition * Direction interaction to the log-transformed, trimmed MI values. We then computed estimated marginal means precisely as described in the previous section.

Results

Speech tracking

We computed MI between the Hilbert-transformed EEG time series and the Hilbert-transformed speech envelopes within three frequency bands of interest that corresponded to the occurrence rates of phrases (0.8–1.1 Hz), words (1.9–2.8 Hz), and syllables (3.5–5.0 Hz) in a cluster of central electrodes.

Specifically, we designed our experiment to assess whether the brain response is driven by the (quasi-)periodic temporal occurrence of linguistic structures and prosody, or whether it is modulated as a function of the linguistic content of those structures. Using MI allowed us to quantify and compare the degree of speech tracking across sentences, word lists, and jabberwocky items.

Our analyses revealed condition-dependent enhanced MI at distinct timescales for the forward conditions (Fig. 4). In the phrase frequency band (0.8–1.1 Hz), the mixed-effects model revealed a significant effect of Condition (Sentence = treatment level; Jabberwocky: $\beta = -0.452$, SE = 0.096, $p < 0.001$; Wordlist: $\beta = -0.491$, SE = 0.116, $p < 0.001$) and Direction (Forward = treatment level; Backward: $\beta = -0.885$, SE = 0.117, $p < 0.001$), as well as Condition * Direction interactions (Jabberwocky * Backward: $\beta = 0.429$, SE = 0.152, $p = 0.008$; Wordlist * Backward: $\beta = 0.523$, SE = 0.185, $p = 0.009$). The estimated marginal means corroborated these results, revealing significant pairwise effects only between the Forward conditions (Sentence–Jabberwocky: $\Delta = 0.452$, SE = 0.098, $p < 0.001$; Sentence–Wordlist: $\Delta = 0.491$, SE = 0.118, $p < 0.001$; all results corrected with Tukey's test for multiple comparisons), but not the backward controls. The observation that

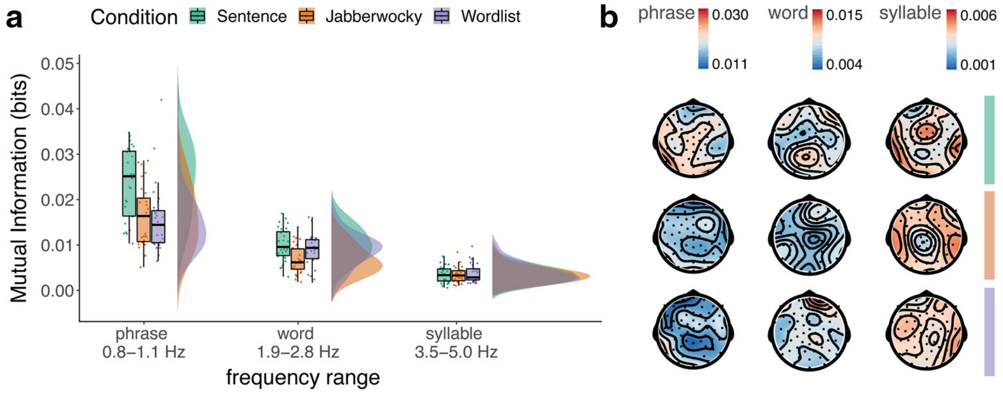

Figure 4. MI between speech signal and brain response. ***a***, MI for Sentences (green), Jabberwocky items (orange), and Wordlists (purple) for phrase, word, and syllable timescales across central electrodes (each dot represents one participant's mean MI response averaged across electrodes). ***b***, Average scalp distribution of MI per condition and band, averaged across participants. Raincloud plots were made using the Raincloud package in R (Allen et al., 2019).

Table 2. Estimated marginal means for mutual information (log-transformed) in the phrase and word frequency bands over a subset of central electrodes

Contrast	Estimate	SE	df	*t*	*p*
Phrase frequency band					
Direction = Forward					
Sentence–Jabberwocky	0.45	0.10	30.0	4.61	<0.01
Sentence–Wordlist	0.49	0.12	30.0	4.17	<0.01
Jabberwocky–Wordlist	0.04	0.10	30.1	0.38	0.93
Direction = Backward					
Sentence–Jabberwocky	0.02	0.11	30.0	0.20	0.98
Sentence–Wordlist	−0.03	0.14	30.0	−0.23	0.97
Jabberwocky–Wordlist	−0.06	0.14	30.0	−0.40	0.92
Word frequency band					
Direction = Forward					
Sentence–Jabberwocky	0.48	0.12	30.0	3.94	<0.01
Sentence–Wordlist	0.16	0.08	29.9	1.96	0.14
Jabberwocky–Wordlist	−0.33	0.14	30.0	−2.40	0.06
Direction = Backward					
Sentence–Jabberwocky	0.25	0.11	30.1	2.31	0.07
Sentence–Wordlist	0.08	0.12	30.1	0.62	0.81
Jabberwocky–Wordlist	−0.18	0.10	30.0	−1.72	0.21

p-value adjustment: Tukey's method for comparing a family of three estimates. Note that pairwise contrasts for the syllable band are omitted here because the linear mixed-effects model showed no significant differences between conditions.

none of the effects was present in the backward speech controls demonstrates that they were not driven by the acoustic properties of the stimuli (Table 2).

In the word frequency band (1.9–2.8 Hz), the mixed-effects model revealed a significant effect of Condition (Sentence = treatment level; Jabberwocky: $\beta = -0.484$, SE = 0.121, $p < 0.001$) and Direction (Forward = treatment level; Backward: $\beta = -0.499$, SE = 0.136, $p < 0.001$). The pairwise contrasts further revealed that this Sentence–Jabberwocky difference was only significant for the forward conditions ($\Delta = 0.484$, SE = 0.123, $p = 0.001$), not for the backward controls (Table 2). Again, this finding indicates that the differences we observed were not driven by differences in the acoustic signals themselves.

In the syllable frequency range (3.5–5.0 Hz), the mixed-effects model revealed no significant effects of Condition (Sentence = treatment level; Jabberwocky: $\beta = 0.001$, SE = 0.121, $p = 0.994$; Wordlist: $\beta = 0.104$, SE = 0.109, $p = 0.348$) or Direction (Forward = treatment level; Backward: $\beta = 0.034$, SE = 0.120, $p = 0.779$), and no interaction between the two (Jabberwocky $*$ Backward: $\beta = 0.144$, SE = 0.166, $p = 0.392$; Wordlist $*$ Backward: $\beta = -0.069$, SE = 0.144, $p = 0.637$).

Together, these findings indicate that neural tracking is enhanced for linguistic structures at timescales specific to the role of that structure in the unfolding meaning of the sentence, consistent with neurophysiologically inspired models of language comprehension (Martin, 2016, 2020; Martin and Doumas, 2017).

An almost identical pattern of results emerged when computing MI over all electrodes (rather than a cluster of central ones). In the phrase frequency range, the mixed-effects model revealed significant effects of Condition (Jabberwocky: $\beta = -0.401$, SE = 0.075, $p < 0.001$; Wordlist: $\beta = -0.418$, SE = 0.088, $p < 0.001$) and Direction (Backward: $\beta = -0.743$, SE = 0.087, $p < 0.001$), as well as significant Condition $*$ Direction interactions (Jabberwocky $*$ Backward: $\beta = 0.296$, SE = 0.099, $p = 0.006$; Wordlist $*$ Backward: $\beta = 0.332$, SE = 0.134, $p = 0.019$). In the word frequency range, the model revealed significant effects of Condition (Jabberwocky: $\beta = -0.407$, SE = 0.093, $p < 0.001$; Wordlist: $\beta = -0.179$, SE = 0.052, $p = 0.002$) and Direction ($\beta = -0.316$, SE = 0.090, $p = 0.002$), but not their interaction. For the forward conditions, the pairwise comparisons further confirmed significantly higher MI for sentences compared with jabberwocky items (Sentence Forward–Jabberwocky Forward: $\Delta = 0.407$, SE = 0.095, $p < 0.001$) and sentences compared with word lists (Sentence Forward–Wordlist Forward: $\Delta = 0.179$, SE = 0.053, $p = 0.006$). Surprisingly, we also found significantly enhanced MI for sentences compared with jabberwocky items in the backward conditions in the word frequency (Sentence Backward–Jabberwocky Backward: $\Delta = 0.288$, SE = 0.083, $p = 0.005$), so we cannot exclude the possibility that this effect is driven to some extent by differences in the acoustic signal. Note, however, that the estimate of this effect is smaller for the backward than the forward differences.

Again, there were no significant effects in the syllable frequency range when computing MI over all electrodes (Condition: Sentence = treatment level; Jabberwocky: $\beta = 0.035$, SE = 0.106, $p = 0.743$; Wordlist: $\beta = 0.037$, SE = 0.090, $p = 0.684$; Direction: Forward = treatment level; Backward: $\beta = 0.147$, SE = 0.124, $p = 0.246$; Jabberwocky $*$ Backward: $\beta = -0.023$, SE = 0.131, $p = 0.859$; Wordlist $*$ Backward: $\beta = -0.045$, SE = 0.130, $p = 0.733$).

Phase MI

When computing MI on the isolated phase values from the Hilbert transform, we again found condition-dependent differences at distinct timescales (Fig. 5, Table 3).

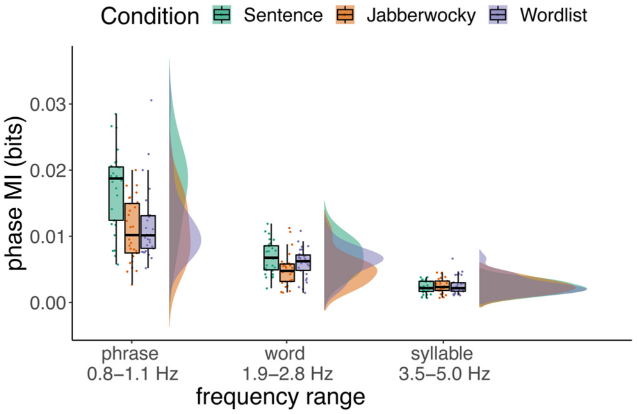

Figure 5. MI between the isolated phase of speech signals and brain responses for Sentences (green), Jabberwocky items (orange), and Wordlists (purple) for phrase, word and syllable timescales across central electrodes (each dot represents one participant's mean MI response averaged across electrodes).

Table 3. Estimated marginal means for phase MI (log-transformed) in the phrase and word frequency bands over a subset of central electrodes

Contrast	Estimate	SE	df	*t*	*p*
Phrase frequency band					
Direction = Forward					
Sentence–Jabberwocky	0.50	0.10	30.0	5.05	<0.01
Sentence–Wordlist	0.40	0.12	30.0	3.36	<0.01
Jabberwocky–Wordlist	−0.10	0.10	30.1	−0.93	0.63
Direction = Backward					
Sentence–Jabberwocky	0.13	0.12	30.0	1.06	0.55
Sentence–Wordlist	0.07	0.13	30.0	0.51	0.87
Jabberwocky–Wordlist	−0.06	0.14	30.0	−0.47	0.89
Word frequency band					
Direction = Forward					
Sentence–Jabberwocky	0.38	0.12	30.1	3.09	0.01
Sentence–Wordlist	0.12	0.07	29.7	1.63	0.25
Jabberwocky–Wordlist	−0.26	0.14	30.1	−1.86	0.17
Direction = Backward					
Sentence–Jabberwocky	0.21	0.12	30.0	1.73	0.21
Sentence–Wordlist	0.06	0.12	30.0	0.54	0.85
Jabberwocky–Wordlist	−0.14	0.10	30.0	−1.42	0.35

p-value adjustment: Tukey's method for comparing a family of three estimates. Note that pairwise contrasts for the syllable band are omitted here because the linear mixed-effects model showed no significant differences between conditions.

In the phrase frequency band (0.8–1.1 Hz), the models revealed significant effects of Condition (Sentence = treatment level; Jabberwocky: $\beta = -0.497$, SE = 0.097, $p < 0.001$; Wordlist: $\beta = -0.402$, SE = 0.118, $p = 0.002$) and Direction ($\beta = -0.805$, SE = 0.106, $p < 0.001$), as well as their interaction (Jabberwocky * Backward: $\beta = 0.368$, SE = 0.150, $p = 0.020$). For the forward conditions, the pairwise contrasts further corroborated these results, with sentences eliciting higher phase MI than jabberwocky items (Sentence Forward–Jabberwocky Forward: $\Delta = 0.497$, SE = 0.099, $p < 0.001$) and sentences eliciting higher phase MI than word lists (Sentence Forward–Wordlist Forward: $\Delta = 0.402$, SE = 0.120, $p = 0.006$; again, all results were corrected by Tukey's test for multiple comparisons).

In the word frequency band (1.9–2.8 Hz), the mixed-effects model revealed a significant effect of Condition (Sentence = treatment level; Jabberwocky: $\beta = -0.380$, SE = 0.121, $p = 0.004$) and Direction ($\beta = -0.474$, SE = 0.126, $p < 0.001$). The pairwise contrasts further revealed significantly higher MI for forward sentences compared with forward jabberwocky items (Sentence Forward–Jabberwocky Forward: $\Delta = 0.380$, SE = 0.123, $p = 0.012$), but not their backward controls. Again, this result demonstrates that the effect is not driven by the acoustic properties of the stimuli (see Table 3 for all pairwise contrasts).

Computing phase MI over all electrodes (rather than a cluster of central ones) revealed a similar pattern of results. In the phrase frequency range, the mixed model revealed significant effects of Condition (Sentence = treatment level; Jabberwocky: $\beta = -0.356$, SE = 0.075, $p < 0.001$; Wordlist: $\beta = -0.309$, SE = 0.089, $p = 0.002$), Direction (Forward = treatment level; Backward: $\beta = -0.662$, SE = 0.076, $p < 0.001$), and their interaction (Jabberwocky * Backward: $\beta = 0.185$, SE = 0.089, $p = 0.047$.) The estimated marginal means showed significant pairwise comparisons only in forward conditions, with forward sentences showing higher phase MI than forward jabberwocky items and forward word lists (Sentence Forward–Jabberwocky Forward: $\Delta = 0.356$, SE = 0.076, $p < 0.001$; Sentence Forward–Wordlist Forward: $\Delta = 0.309$, SE = 0.091, $p = 0.005$), and no significant effects for the backward comparisons (Sentence Backward–Jabberwocky Backward: $\Delta = 0.171$, SE = 0.102, $p = 0.227$; Sentence Backward–Wordlist Backward: $\Delta = 0.099$, SE = 0.110, $p = 0.644$; Jabberwocky Backward–Wordlist Backward: $\Delta = -0.072$, SE = 0.125, $p = 0.833$).

In the word frequency band, the mixed-effects model revealed significant effects of Condition (Sentence = treatment level; Jabberwocky: $\beta = -0.329$, SE = 0.089, $p < 0.001$; Wordlist: $\beta = -0.139$, SE = 0.045, $p = 0.005$) and Direction ($\beta = -0.351$, SE = 0.091, $p < 0.001$). The estimated marginal means further corroborated this finding only in the forward conditions (Sentence Forward–Jabberwocky Forward: $\Delta = 0.329$, SE = 0.091, $p = 0.003$; Sentence Forward–Wordlist Forward: $\Delta = 0.139$, SE = 0.046, $p = 0.014$). In contrast to the "general" MI values, we found no significant differences between the backward controls when computing the isolated phase MI over the entire head (Sentence Backward–Jabberwocky Backward: $\Delta = 0.209$, SE = 0.101, $p = 0.112$; Sentence Backward–Wordlist Backward: $\Delta = 0.088$, SE = 0.087, $p = 0.577$; Jabberwocky Backward–Wordlist Backward: $\Delta = -0.121$, SE = 0.085, $p = 0.343$). Again, these findings are consistent with neurophysiologically inspired models of language comprehension (Martin, 2016, 2020; Martin and Doumas, 2017).

Tracking of abstract linguistic units

Inspecting the modulation spectra of our stimuli (Fig. 2), it is apparent that—although carefully designed—the acoustic signals are not entirely indistinguishable between conditions based on their spectral properties. Most notably, Sentence stimuli appear to exhibit a small peak at ~0.5 Hz (roughly corresponding to the phrase timescale in our stimuli) compared with the other two conditions. It is important to note that (1) differences between conditions are not surprising, given that our stimuli were naturally spoken; and (2) we specifically designed our experiment to include backward versions of all conditions to control for slight differences between the acoustic envelopes of the forward stimuli. That being said, we conducted an additional, exploratory analysis to further reduce the potential confound of differences between the acoustic modulation spectra and to disentangle the distribution of linguistic phrase representations and the acoustic stimulus even further. Specifically, we computed MI in the delta–theta range (0.8–5 Hz) between the brain response and abstracted dimensionality-reduced annotations of all stimuli, containing only information about when words could be integrated into phrases (Brodbeck et al., 2018; see Materials and Methods for detailed descriptions of how these annotations were created).

Table 4. Estimated marginal means for MI (log-transformed) calculated over abstract phrase representations

Contrast	Estimate	SE	df	t	p
Direction = Forward					
Sentence–Jabberwocky	0.33	0.11	29.8	2.85	0.02
Sentence–Wordlist	0.52	0.12	30.0	4.25	<0.01
Jabberwocky–Wordlist	0.20	0.13	30.0	1.54	0.29
Direction = Backward					
Sentence–Jabberwocky	−0.03	0.12	30.0	−0.22	0.97
Sentence–Wordlist	−0.10	0.10	30.1	−0.99	0.59
Jabberwocky–Wordlist	−0.07	0.11	30.1	−0.69	0.77

p-value adjustment: Tukey's method for comparing a family of three estimates.

These annotation-based analyses revealed significant effects of Condition (Sentence = treatment level; Jabberwocky: $\beta = -0.326$, SE = 0.112, $p = 0.007$; Wordlist: $\beta = -0.521$, SE = 0.120, $p < 0.001$), Direction ($\beta = -0.754$, SE = 0.115, $p < 0.001$), and their interaction (Jabberwocky * Backward: $\beta = 0.352$, SE = 0.164, $p = 0.040$; Wordlist * Backward: $\beta = 0.621$, SE = 0.156, $p < 0.001$). The estimated marginal means further revealed increased MI for forward sentences compared with forward jabberwocky items and forward word lists (Sentence Forward–Jabberwocky Forward: $\Delta = 0.326$, SE = 0.114, $p = 0.021$; Sentence Forward–Wordlist Forward: $\Delta = 0.521$, SE = 0.123, $p < 0.001$; all results were corrected with Tukey's for multiple comparisons) and no significant difference among the backward controls (Table 4).

Again, the same pattern of results also emerged when computing MI over all electrodes: the mixed-effects model revealed significant effects of Condition (Jabberwocky: $\beta = -0.365$, SE = 0.087, $p < 0.001$; Wordlist: $\beta = -0.611$, SE = 0.098, $p < 0.001$), Direction ($\beta = -0.813$, SE = 0.090, $p < 0.001$), and their interaction (Jabberwocky * Backward: $\beta = 0.390$, SE = 0.148, $p = 0.014$; Wordlist * Backward: $\beta = 0.678$, SE = 0.131, $p < 0.001$). The pairwise contrasts were, again, only significant between the forward conditions (Sentence Forward–Jabberwocky Forward: $\Delta = 0.365$, SE = 0.088, $p < 0.001$; Sentence Forward–Wordlist Forward: $\Delta = 0.611$, SE = 0.100, $p < 0.001$), but not the backward controls (Sentence Backward–Jabberwocky Backward: $\Delta = -0.024$, SE = 0.109, $p = 0.973$; Sentence Backward–Wordlist Backward: $\Delta = -0.067$, SE = 0.097, $p = 0.770$; Jabberwocky Backward–Wordlist Backward: $\Delta = -0.043$, SE = 0.100, $p = 0.905$). These results support our previously reported findings, showing that neural tracking is influenced by the presence of abstract linguistic information. In other words, this exploratory analysis supports our earlier finding that the "sensitivity" of the brain to linguistic structure and meaning goes above and beyond the acoustic signal and both word-level semantic and prosodic controls.

Discussion

The current experiment tested how the brain attunes to linguistic information. Contrasting sentences, word lists and jabberwocky items, we analyzed, by proxy, how the brain response is modulated by sentence-level prosody, lexical semantics, and compositional structure and meaning. Our findings show that (1) the neural response is driven by compositional structure and meaning, beyond both acoustic-prosodic and lexical information; and (2) the brain most closely tracks the most structured representations on the timescales we analyzed. To our knowledge, this is the first study to systematically disentangle the contribution of linguistic content from its timing and rhythm in natural speech by using linguistically informed controls. Additionally, our data demonstrate cortical tracking of naturalistic language without a nonlinguistic task such as syllable counting and outlier trial or target-detection tasks. We show that oscillatory activity attunes to structured and meaningful content, suggesting that neural tracking reflects computations related to inferring linguistic representations from speech, and not merely tracking of rhythmicity or timing. We discuss these findings in more detail below.

Using mutual information analysis, we quantified the degree of speech tracking in frequency bands corresponding to the timescales at which linguistic structures (phrases, words, and syllables) could be inferred from our stimuli. On the phrase timescale, we found that sentences had the most shared information between stimulus and response. Crucially, this is not merely a chunking mechanism (Bonhage et al., 2017; Ghitza, 2017)—participants could have "chunked" the word lists (which have their own naturally produced nonsentential prosody) into units of adjacent words, and the jabberwocky items into prosodic units. This is especially interesting given recent work by Jin et al. (2020), showing that enhanced delta-band activity can be "induced" in listeners by teaching them to chunk a sequence of (synthesized) words according to different sets of artificial grammar rules. Conversely, the observed patterns of activity cannot exclusively be driven by the lexico-semantic content of our stimuli (Frank and Yang, 2018)—sentences and word lists contained the same lexical items, yet MI was enhanced for Sentence stimuli, where words could be combined into phrases and higher-level representations. As such, we argue that the dominating process we observe appears to be processing compositional semantic structure, above and beyond prosodic chunking and word-level meaning. We show that the brain aligns more to periodically occurring units when they contain meaningful information and are thus relevant for linguistic processing.

On the word timescale, the emerging picture is somewhat more diverse than on the phrase timescale. Specifically, we found enhanced tracking for sentences compared with jabberwocky items. We tentatively take this finding to indicate that, at the word timescale, the dominant process appears to be context-dependent word recognition—perhaps based in perceptual inference. This is further corroborated by the results of computing MI over all electrodes, rather than a subset, with sentences eliciting higher MI than both jabberwocky items and word lists. Note, however, that we also found enhanced MI on the word timescale for word lists compared with jabberwocky items in the backward controls when computing MI over all electrodes. Here, listeners could not have processed words within the context of phrases or sentences, which makes it somewhat difficult to integrate these results. One possible explanation for this surprising finding might be that there is still some acoustic-prosodic information available in the backward controls that distinguishes word lists from jabberwocky items. Future research could address this in detail, for example by including a control condition with entirely flat prosody (Ding et al., 2016, 2017a).

There continues to be a vibrant debate about whether language-related cortical activity in the delta–theta range is truly oscillatory in nature or whether the observed patterns of neural activity arise as a series of evoked responses (Haegens and Golumbic, 2018; Rimmele et al., 2018; Zoefel et al., 2018; Obleser and Kayser, 2019). Our current results cannot speak to this question; in fact, we have been careful to refer to our results as "tracking" rather than "entrainment" throughout this article. To be clear, we do not take the observed increased MI for sentences compared with jabberwocky items and word lists as evidence for an intrinsic "phrase-level oscillator" or "word-level oscillator." Rather, we interpret our findings as a manifestation of the cortical computations that may occur during language comprehension. Here, we observe them in the delta frequency range because that is the timescale on which higher-level linguistic units occur in our stimuli.

Many previous studies have shown that attention can modulate neural entrainment (Haegens et al., 2011; Ding and Simon, 2012; Golumbic et al., 2013; Lakatos et al., 2013; Calderone et al., 2014). Importantly, Ding et al. (2018) found that tracking beyond the syllable envelope requires attention to the speech stimulus. In our current experiment, participants were instructed to attentively listen to the audio recordings in all conditions, but it is possible that “attending to sentences” might be easier than “attending to jabberwocky items,” and that listeners pay closer attention to higher-level structures in intelligible and meaningful speech. Additionally, our study used a block design, which could, in principle, have encouraged participants to use different attentional resources during the different blocks. As such, we cannot rule out the possibility that our effects might be influenced by a mechanism based on attentional control. It is, however, difficult to disentangle “attention” from “comprehension” in this kind of argument: meaningful information within a stimulus can arguably only lead to increased attention if it is comprehensible. We plan to investigate these questions in future experiments.

Overall, the pattern of results is consistent with cue integration-based models of language processing (Martin, 2016, 2020), where the activation profile of different populations of neurons over time encodes linguistic structure as it is inferred from sensory correlates in real time (Martin and Doumas, 2017). The model of language processing of Martin (2016, 2020) builds on and extends neurophysiological models of cue integration, where percepts are inferred from sensory cues through summation and normalization, both of which have been proposed as canonical neural computations (Carandini and Heeger, 1994, 2011; Ernst and Bülthoff, 2004; Landy et al., 2011; Fetsch et al., 2013; for cue integration-based models of speech and word recognition, see Norris and McQueen, 2008; Toscano and McMurray, 2010; McMurray and Jongman, 2011). Martin (2016, 2020) proposed that, during all stages of language processing, the brain might draw on these same neurophysiological computations.

Crucially, inferring linguistic representations from speech sounds requires not only bottom-up sensory information, but also top-down memory-based cues (Marslen-Wilson, 1987; Kaufeld et al., 2020). Martin (2016, 2020) therefore suggested that cue integration during language comprehension is an iterative process, where cues that have been inferred from the acoustic signal can, in turn, become cues for higher levels of processing. The pattern of findings in our current experiment strongly speaks to cue integration-based models of language comprehension: we observe that tracking of the speech signal is enhanced when meaningful linguistic units can be inferred, suggesting that the alignment of populations of neurons might, indeed, encode the generation of inference-based linguistic representations (Martin and Doumas, 2017).

Our results also speak to analysis-by-synthesis-based accounts of speech processing, more generally (Halle and Stevens, 1962; Bever and Poeppel, 2010; Poeppel and Monahan, 2011). In an analysis-by-synthesis model of speech perception, speech recognition is achieved by internally generating (synthesizing) patterns according to internal rules, and matching (analyzing) them against the acoustic input signal. Similarly, our findings are in line with the notion of hierarchical temporal receptive windows from early sensory to higher-level perceptual and cognitive brain areas (Hasson et al., 2008; Lerner et al., 2011).

There are, of course, many open questions that arise from our results. Perhaps most obviously (although presumably limited by the resolution of time–frequency analysis), it would be interesting to investigate how “far” cue integration can be traced during even more natural language comprehension situations (Alday, 2019; Alexandrou et al., 2020). To what degree are higher-level linguistic cues, such as sentential, contextual, or pragmatic information, encoded in the neural response? Another interesting avenue for future research would be to investigate whether similar patterns can be observed during language production. Martin (2016, 2020) suggested that not only language comprehension, but also language production draws on principles of cue integration. Finally—and consequentially, if cue integration underlies both comprehension and production processes—we would be curious to learn more about cue integration “in action,” specifically during dialogue settings, where interlocutors comprehend and plan utterances nearly simultaneously.

In summary, this study showed that speech tracking is sensitive to linguistic structure and meaning, above and beyond prosodic and lexical-semantic controls. In other words, content determines tracking, not just timescale. This extends previous findings and advances our understanding of spoken language comprehension in general, because our experimental manipulation allows us, for the first time, to disentangle the influence of linguistic structure and meaning on the neural response from word-level meaning and prosodic regularities occurring in naturalistic stimuli.

References

Alday PM (2019) M/EEG analysis of naturalistic stories: a review from speech to language processing. Lang Cogn Neurosci 34:457–473.

Alexandrou AM, Saarinen T, Kujala J, Salmelin R (2020) Cortical entrainment: what we can learn from studying naturalistic speech perception. Lang Cogn Neurosci 35:681–693.

Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit RA (2019) Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res 4:63.

Audacity Team (2014) Audacity(R): free audio editor and recorder, version 2.4.1. [computer application].

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Soft 67:1–48.

Boersma P, Weenink D (2020) Praat: doing phonetics by computer version 6.1.15. Amsterdam: Phonetic Sciences, University of Amsterdam.

Bever TG, Poeppel D (2010) Analysis by synthesis: a (re-) emerging program of research for language and vision. Biolinguistics 4:174–200.

Bonhage CE, Meyer L, Gruber T, Friederici AD, Mueller JL (2017) Oscillatory EEG dynamics underlying automatic chunking during sentence processing. Neuroimage 152:647–657.

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.

Calderone DJ, Lakatos P, Butler PD, Castellanos FX (2014) Entrainment of neural oscillations as a modifiable substrate of attention. Trends Cogn Sci 18:300–309.

Carandini M, Heeger DJ (1994) Summation and division by neurons in primate visual cortex. Science 264:1333–1336.

Carandini M, Heeger DJ (2011) Normalization as a canonical neural computation. Nat Rev Neurosci 13:51–62.

Cogan GB, Poeppel D (2011) A mutual information analysis of neural coding of speech by low-frequency MEG phase information. J Neurophysiol 106:554–563.

Ding N, Simon JZ (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107: 78–89.

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, Melloni L, Yang A, Wang Y, Zhang W, Poeppel D (2017a) Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). Front Hum Neurosci 11:481.

Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D (2017b) Temporal modulations in speech and music. Neurosci Biobehav Rev 81:181–187.

Ding N, Pan X, Luo C, Su N, Zhang W, Zhang J (2018) Attention is required for knowledge-based sequential grouping: insights from the integration of syllables into words. J Neurosci 38:1178–1188.

Doelling KB, Arnal LH, Ghitza O, Poeppel D (2014) Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage 85:761–768.

Eisner F, McQueen JM (2018) Speech perception. In: Stevens' handbook of experimental psychology and cognitive neuroscience, Ed 4 (Stevens SS, Wixted JT, eds), pp 1–46. New York: Wiley.

Ernst MO, Bülthoff HH (2004) Merging the senses into a robust percept. Trends Cognitive Sci 8:162–169.

Fetsch CR, DeAngelis GC, Angelaki DE (2013) Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. Nat Rev Neurosci 14:429–442.

Frank SL, Yang J (2018) Lexical representation explains cortical entrainment during speech comprehension. PLoS One 13:e0197304.

Ghitza O (2017) Acoustic-driven delta rhythms as prosodic markers. Lang Cogn Neurosci 32:545–561.

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

Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P, Garrod S (2013) Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol 11:e1001752.

Haegens S, Golumbic EZ (2018) Rhythmic facilitation of sensory processing: a critical review. Neurosci Biobehav Rev 86:150–165.

Haegens S, Händel BF, Jensen O (2011) Top-down controlled alpha band activity in somatosensory areas determines behavioral performance in a discrimination task. J Neurosci 31:5197–5204.

Halle M, Stevens K (1962) Speech recognition: A model and a program for research. IRE transactions on information theory 8:155–159.

Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N (2008) A hierarchy of temporal receptive windows in human cortex. J Neurosci 28:2539–2550.

Ince RA, Giordano BL, Kayser C, Rousselet GA, Gross J, Schyns PG (2017) A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula. Hum Brain Mapp 38:1541–1573.

Jin P, Lu Y, Ding N (2020) Low-frequency neural activity reflects rule-based chunking during speech listening. Elife 9:e55613.

Kaufeld G, Ravenschlag A, Meyer AS, Martin AE, Bosker HR (2020) Knowledge-based and signal-based cues are weighted flexibly during spoken language comprehension. J Exp Psychol Learn Mem Cogn 46:549–562.

Kayser SJ, Ince RAA, Gross J, Kayser C (2015) Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. J Neurosci 35:14691–14701.

Keitel A, Ince RAA, Gross J, Kayser C (2017) Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. Neuroimage 147:32–42.

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

Keuleers E, Brysbaert M (2010) Wuggy: a multilingual pseudoword generator. Behav Res Methods 42:627–633.

Kösem A, Bosker HR, Takashima A, Meyer A, Jensen O, Hagoort P (2018) Neural entrainment determines the words we hear. Curr Biol 28:2867–2875.

Kutas M, Federmeier KD (2000) Electrophysiology reveals semantic memory use in language comprehension. Trends Cogn Sci 4:463–470.

Kutas M, Van Petten CK, Kluender R (2006) Psycholinguistics electrified II (1994–2005). In: Handbook of psycholinguistics (Gernsbacher MA, Traxler MJ, eds), pp 659–724. Amsterdam: Academic.

Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest Package: tests in linear mixed effects models. J Stat Soft 82:1–26.

Lakatos P, Musacchia G, O'Connel MN, Falchier AY, Javitt DC, Schroeder CE (2013) The spectrotemporal filter mechanism of auditory selective attention. Neuron 77:750–761.

Landy MS, Banks MS, Knill DC (2011) Ideal-observer models of cue integration. In: Sensory cue integration (Trommershäuser J, Kording K, Landy MS, eds), pp 5–29. Oxford, UK: Oxford UP.

Length R (2018) emmeans: estimated marginal means, aka least-square means. R package version 1.3.0. Vienna, Austria: R Foundation.

Lerner Y, Honey CJ, Silbert LJ, Hasson U (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J Neurosci 31:2906–2915.

Marslen-Wilson WD (1987) Functional parallelism in spoken word-recognition. Cognition 25:71–102.

Martin AE (2016) Language processing as cue integration: grounding the psychology of language in perception and neurophysiology. Front Psychol 7:120.

Martin AE (2020) A compositional neural architecture for language. J Cogn Neurosci 32:1407–1427.

Martin AE, Doumas LAA (2017) A mechanism for the cortical computation of hierarchical linguistic structure. PLoS Biol 15:e2000663.

McMurray B, Jongman A (2011) What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychol Rev 118:219–246.

Meyer L, Sun Y, Martin AE (2019) Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. Lang Cogn Neurosci 35:1089–1099.

Norris D, McQueen JM (2008) Shortlist B: a Bayesian model of continuous speech recognition. Psychol Rev 115:357–395.

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

Obleser J, Herrmann B, Henry MJ (2012) Neural oscillations in speech: don't be enslaved by the envelope. Front Hum Neurosci 6:250.

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 2011:156869.

Peelle JE, Davis MH (2012) Neural oscillations carry speech rhythm through to comprehension. Front Psychol 3:320.

Poeppel D, Monahan PJ (2011) Feedforward and feedback in speech perception: Revisiting analysis by synthesis. Language and Cognitive Processes 26:935–951.

R Core Team (2018) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rimmele JM, Morillon B, Poeppel D, Arnal LH (2018) Proactive sensing of periodic and aperiodic auditory patterns. Trends Cogn Sci 22:870–882.

Smith ZM, Delgutte B, Oxenham AJ (2002) Chimaeric sounds reveal dichotomies in auditory perception. Nature 416:87–90.

Soderstrom M, Seidl A, Nelson DGK, Jusczyk PW (2003) The prosodic bootstrapping of phrases: evidence from prelinguistic infants. J Mem Lang 49:249–267.

Toscano JC, McMurray B (2010) Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cogn Sci 34:434–464.

Zoefel B, VanRullen R (2015) The role of high-level processes for oscillatory phase entrainment to speech sound. Front Hum Neurosci 9:651.

Zoefel B, ten Oever S, Sack AT (2018) The involvement of endogenous neural oscillations in the processing of rhythmic input: more than a regular repetition of evoked neural responses. Front Neurosci 12:95.