Auditory word comprehension is less incremental in isolated words

Phoebe Gaston,[1,^,*] Christian Brodbeck,[2,^] Colin Phillips,[1] and Ellen Lau[1]

[1]Department of Linguistics, University of Maryland, College Park, MD 20742
[2]Institute for Systems Research, University of Maryland, College Park, MD 20742

[^]Denotes equal contribution. Phoebe Gaston and Christian Brodbeck are now at Department of Psychological Sciences, University of Connecticut, Storrs, CT 06269.

* Corresponding author: phoebe.gaston@uconn.edu

Acknowledgements

We thank Daphne Amir, Fen Ingram, and Stephanie Pomrenke for assistance with stimulus selection, and Aura Cruz Heredia for assistance with some of the data collection.

Conflict of Interest

Authors report no conflict of interest.

Funding Sources

This material is based upon work supported by the National Science Foundation under Grants BCS-1749407 (E. Lau, PI) and DGE-1449815 (C. Phillips, PI) at the University of Maryland. Phoebe Gaston was also supported by a Flagship Fellowship from the University of Maryland and by NIH T32 DC017703 (I-M Eigsti & E. Myers, PIs) at the University of Connecticut.

Auditory word comprehension is less incremental in isolated words

Abstract

Speech input is often understood to trigger rapid and automatic activation of successively higher-level representations for comprehension of words. Here we show evidence from magnetoencephalography that incremental processing of speech input is limited when words are heard in isolation as compared to continuous speech. This suggests a less unified and automatic process than is often assumed. We present evidence that neural effects of phoneme-by-phoneme lexical uncertainty, quantified by cohort entropy, occur in connected speech but not isolated words. In contrast, we find robust effects of phoneme probability, quantified by phoneme surprisal, during perception of both connected speech and isolated words. This dissociation rules out models of word recognition in which phoneme surprisal and cohort entropy are common indicators of a uniform process, even though these closely related information-theoretic measures both arise from the probability distribution of wordforms consistent with the input. We propose that phoneme surprisal effects reflect automatic access of a lower level of representation of the auditory input (e.g., wordforms) while cohort entropy effects are task-sensitive, driven by a competition process or a higher-level representation that is engaged late (or not at all) during the processing of single words.

Introduction

Speech recognition necessarily involves the access of multiple levels of representation in response to auditory input, from phonemes to wordforms to higher-level lexical-syntactic representations that link wordforms to meaning. While much about this process remains to be elucidated, research on spoken word recognition has reached broad consensus on several points. The contributions of a vast behavioral literature (reviewed by, e.g., Dahan & Magnuson (2006); McQueen (2007); Magnuson, Mirman, & Myers (2013); Magnuson (2016)) indicate an incremental, phoneme-by-phoneme process of winnowing down the phonological wordforms that are consistent with the unfolding auditory input (e.g., Grosjean (1980); Zwitserlood (1989); Allopenna et al. (1998); and following). Conceptual information associated with those wordforms can be incrementally activated (e.g., Zwitserlood (1989); Yee & Sedivy (2006); and following), and syntactic information is rapidly invoked (e.g., Marslen-Wilson & Tyler (1980); McAllister (1988); and following). This process is highly sensitive to distributional statistics, captured by word frequency (e.g., Connine et al. (1990); Dahan et al. (2001)).

The evidence leading to this consensus comes from a broad array of experimental approaches that vary in which aspects of word recognition they can most effectively probe. These approaches use stimuli that vary from sublexical phoneme sequences to natural, connected speech. Combining evidence from these different paradigms is usually guided by an assumption that there is a uniform, automatic progression of processing triggered by speech input, such that we can expect datapoints from different points in that progression to cohere. Under this assumption, simpler or single-word paradigms will straightforwardly capture the fundamental word recognition sequence in isolation, while presenting more complex input allows us to investigate how contextual information influences, for example, the speed of processing or the set of lexical candidates under consideration.

In **Figure 1**, we sketch a representative sequence of processing proposed to occur in response to each phoneme of speech input. TRACE (McClelland & Elman, 1986) is an example of a model that is consistent with the illustrated principles. Each level of representation automatically determines the most likely interpretation of the input through local competition and broadcasts this interpretation through feed-forward and feed-back connections. The assumption of automaticity implies that any speech input engages this processing hierarchy in the same manner. The task context might change the information available at different levels, but not the basic sequence of processing. However, if the assumption of automaticity is incorrect, then the basic process of word recognition could deviate significantly according to the demands of different comprehension scenarios. This deviation could occur because of variation in, for instance, the relevance of different types of information to different experimental tasks, the ease of word segmentation, and the degree to which word-to-word dependencies occur in the input.

In this paper we present neural evidence that word recognition in isolation may proceed in a qualitatively different way from word recognition in continuous speech. Behavioral measures or paradigms requiring an explicit response to each stimulus make comparison between isolated words and continuous speech difficult, and a single trial generally reflects the status of just a single item in the lexicon. Instead, we turn to a neural measure-- temporal response function (TRF) analysis of magnetoencephalography (MEG) responses—that can be applied in exactly the same way to single-word and continuous-speech listening, and that reflects distributional

properties of the entire class of word candidates consistent with each presented phoneme. We show that the effects of two measures that have both been understood to reflect automatic wordform-level processing in fact dissociate robustly according to the nature of the experiment. This dissociation implicates a break in the automaticity of the sequence of activation and indicates a difference between the processing of words presented in isolation and words presented in continuous speech. Our findings have implications for the architecture of word recognition models as well as for experimental approaches to studying speech perception.

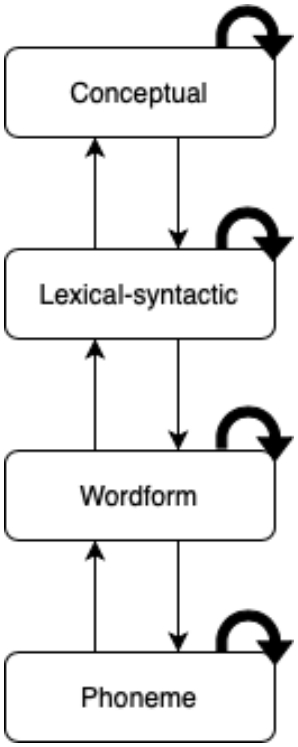

Figure 1. *Automatic sequence of processing assumed to occur in response to each phoneme of speech input. Straight arrows indicate connections between levels of representation. Curved arrows indicate a within-level competition/selection process.*

Phoneme surprisal and cohort entropy

The neural response to speech has been shown to be modulated by information-theoretic properties of the set of wordforms that match the auditory input at any given phoneme (Brodbeck et al., 2018, 2021; Di Liberto et al., 2019; Donhauser & Baillet, 2020; Ettinger et al., 2014; Gagnepain et al., 2012; Gaston & Marantz, 2018; Gillis et al., 2021; Gwilliams et al., 2020; Gwilliams & Marantz, 2015; Kocagoncu et al., 2017). Two of these properties in particular – cohort entropy and phoneme surprisal – have emerged as promising means of investigating the time course of auditory word recognition.

Phoneme surprisal at a given phoneme reflects the conditional probability of that phoneme given the preceding sequence of phonemes in the current word. Phoneme surprisal at position *i* in a wordform is defined as $-log_2\ p(k_i \mid k_1, \ldots k_{i-1})$ where k_i is the phoneme at position *i* and *i* = 1 for the first phoneme in the wordform. Cohort entropy at that same phoneme, in contrast, is determined by the probability distribution over wordforms that might complete that phoneme sequence, reflecting how much uncertainty there is among those candidates. Cohort entropy at

position *i* in a wordform is defined as $-\sum_{w}^{C_i}(p(w \mid k_1, \ldots k_i) \times log_2\ p(w \mid k_1, \ldots k_i))$ where w is each wordform in the cohort C_i of wordforms consistent with the sequence of phonemes $k_1, \ldots k_i$. One of the critical differences between these formulations is that cohort entropy is forward-looking in a way that phoneme surprisal is not. A cohort entropy effect reflects expectations for potential candidates that would be consistent with the current input, while phoneme surprisal provides evidence of the degree to which a phoneme could previously have been expected.

More neural activity is generally observed in response to higher surprisal, or lower probability, phonemes, consistent with many cognitive domains in which predictable or higher probability stimuli elicit reduced neural responses (see Aitchison & Lengyel (2017)). Exactly how cohort entropy should be expected to drive neural activity is less clear. A larger set of candidates has a higher cohort entropy than a smaller set of candidates, and a set of candidates in which probability is equally distributed has a higher cohort entropy than a set of candidates in which probability is concentrated on a single candidate. Greater uncertainty could be associated with more neural activity due to an intensified process of lexical competition (Gagnepain et al., 2012), or due to increased attentional gain on bottom-up input (Donhauser & Baillet, 2020), or it could be that lower uncertainty is a precondition for other processes to be engaged (Ettinger et al., 2014).

Despite these differences, phoneme surprisal and cohort entropy are often investigated and presented in tandem as interchangeable indicators of wordform-level processing. One likely reason for this approach in the literature is that the conditional phoneme probabilities underlying both measures are calculated from the probabilities of wordforms consistent with the input. The two variables are also often correlated, and their effects in neural data frequently co-occur. Finally, in a hypothesized model of word recognition that includes automatic engagement of successive representational levels regardless of task or context, phoneme surprisal and cohort entropy effects are simply two different windows into the same automatic flow of activation through the system.

Variation in neural effects of cohort entropy and phoneme surprisal

Despite frequently being treated interchangeably, a careful look at the prior literature reveals considerable variation in whether, where, and when phoneme surprisal and cohort entropy effects manifest across experiments. This variation has not previously been examined systematically. Thus, before we proceed to our own study, we review this literature and consider whether there are properties of the stimulus or experimental context that can help explain when cohort entropy and phoneme surprisal effects do or do not occur, and what this might mean for the processes and levels of representation they describe. An account of this variability is important for improving the utility of phoneme surprisal and cohort entropy as measures for investigating speech perception and specifically the class of active items in competition for recognition at any given point in a word. However, understanding this variability also has the potential to illuminate dissociable sub-processes in word recognition.

We begin by trying to characterize why these effects occur at all in some experiments and not in others, though further efforts to understand variation in the localization and time course of these effects will also be important. In Table 1, we summarize existing electrophysiology (primarily MEG) studies that have tested for effects of cohort entropy and phoneme surprisal on neural

activity. Effects of both cohort entropy and phoneme surprisal have been reported in behavioral measures of auditory word recognition (Baayen et al., 2007; Balling & Baayen, 2012; Bien et al., 2011; Kemps et al., 2005; Wurm et al., 2006). However, testing for such effects in behavioral data generally requires constructing a cumulative measure of a phoneme-level variable across the course of the word or selecting the variable's value at just one phoneme position as the predictor. Therefore we restrict our focus here to neural measures that have the temporal resolution to examine cohort entropy and phoneme surprisal effects on a phoneme-by-phoneme basis. We exclude one additional study on the processing of continuous speech (Di Liberto et al., 2019), which did not report effects of cohort entropy and phoneme surprisal separately.

Table 1. Properties of the Stimulus and Experimental Task for Existing Electrophysiology Studies Reporting Phoneme Surprisal or Cohort Entropy Effects

Study	Phoneme surprisal effect?	Cohort entropy effect?	Stimulus	Experimental task	Multimorphemic words included?
Gagnepain et al. (2012)	yes	no	single words	pause detection	no
Ettinger et al. (2014)	yes	yes^	single words	lexical decision	yes
Brennan et al. (2014)		no	single words	lexical decision	no
Lewis & Poeppel (2014)		no	single words	lexical decision	no
Gwilliams & Marantz (2015)	yes		single words	lexical decision	yes
Kocagoncu et al. (2017)		yes^	single words	nonword detection	not specified
Gaston & Marantz (2018)	yes	yes*	three-word phrases	phrase acceptability	yes
Brodbeck, Hong, et al. (2018)	yes	yes	continuous speech	comprehension questions	yes
Donhauser & Baillet (2020)	yes		continuous speech	comprehension questions	yes
Gwilliams et al. (2020)	yes	yes	continuous speech	comprehension questions	yes
Gillis et al. (2021)	yes	yes	continuous speech	comprehension questions	yes
Brodbeck et al. (2021)	yes	yes	continuous speech	comprehension questions	yes

Note. Grey cells indicate studies that did not test for the specified effect. Superscripts indicate that a reported cohort entropy effect did not survive when phoneme surprisal was controlled for (*), or that such a test was not performed (^).

Table 1 demonstrates that phoneme surprisal and cohort entropy effects have very different profiles across studies. Phoneme surprisal effects were reported in all studies that tested for them (Brodbeck et al., 2018, 2021; Donhauser & Baillet, 2020; Ettinger et al., 2014; Gagnepain et al., 2012; Gaston & Marantz, 2018; Gillis et al., 2021; Gwilliams et al., 2020; Gwilliams & Marantz, 2015), and thus appear to be robust to variation in stimulus and experimental task. Cohort entropy, in contrast, produces mixed results. Among studies that presented single words and short phrases, three reported cohort entropy effects (Ettinger et al., 2014; Gaston & Marantz, 2018; Kocagoncu et al., 2017) and three tested for but failed to find them (Brennan et al., 2014; Gagnepain et al., 2012; Lewis & Poeppel, 2014). The presence or absence of multimorphemic words in the study is potentially relevant, as the three studies that failed to find cohort entropy effects included only monomorphemic words. However, more important in our view is that the three single-word studies that reported cohort entropy effects did not exclude the possibility that these effects were due to the highly correlated phoneme surprisal measure. Gaston and Marantz (2018) in fact found that their significant cohort entropy effect was no longer significant in a model that controlled for phoneme surprisal, and the other two studies (Ettinger et al., 2014; Kocagoncu et al., 2017) did not conduct such a test. In continuous speech, cohort entropy effects were reported in all studies that tested for them (Brodbeck et al., 2018, 2021; Gillis et al., 2021; Gwilliams et al., 2020), with methods that controlled for effects of phoneme surprisal. We conclude that, in the existing electrophysiology literature, there is strong evidence for phoneme surprisal effects across the board, but for cohort entropy effects only in continuous speech.

A true dissociation between cohort entropy and phoneme surprisal effects would indicate not only that these measures do not index the same level of representation or process, but also that whatever drives cohort entropy effects does not occur during the processing of single words, or at least does not occur incrementally (i.e., phoneme by phoneme). This is not consistent with all processing steps being engaged in a fully automatic sequence during speech recognition. However, this interpretation of the prior literature is complicated by the fact that many of these studies did not control for potential confounds, such as acoustic variables and overlapping responses to different phonemes. Differences in statistical power or analysis methods (which vary widely) may also have contributed to the apparent influence of stimulus on cohort entropy effects.

The current study

Hypothesizing that cohort entropy and phoneme surprisal do, indeed, dissociate, and that cohort entropy effects do not occur for single words, we evaluated cohort entropy and phoneme surprisal effects on the neural response to speech in a simple single-word paradigm and then directly compared this data to an existing continuous-speech dataset (Brodbeck et al., 2021). Comparing single-word and continuous-speech data requires that the two types of responses be evaluated with the same method. Analysis techniques traditionally applied to single-word studies are not suitable for responses to continuous speech, and generally fail to account for acoustic and other confounding variables, as well as the overlapping nature of phoneme responses. Instead, we modeled source-localized MEG data with temporal response functions (**Figure 2**), a method that deals with acoustic confounds and was originally developed for continuous speech. This allowed for novel comparison between single words and continuous speech as well as a more accurate characterization of the single-word response relative to previous analyses.

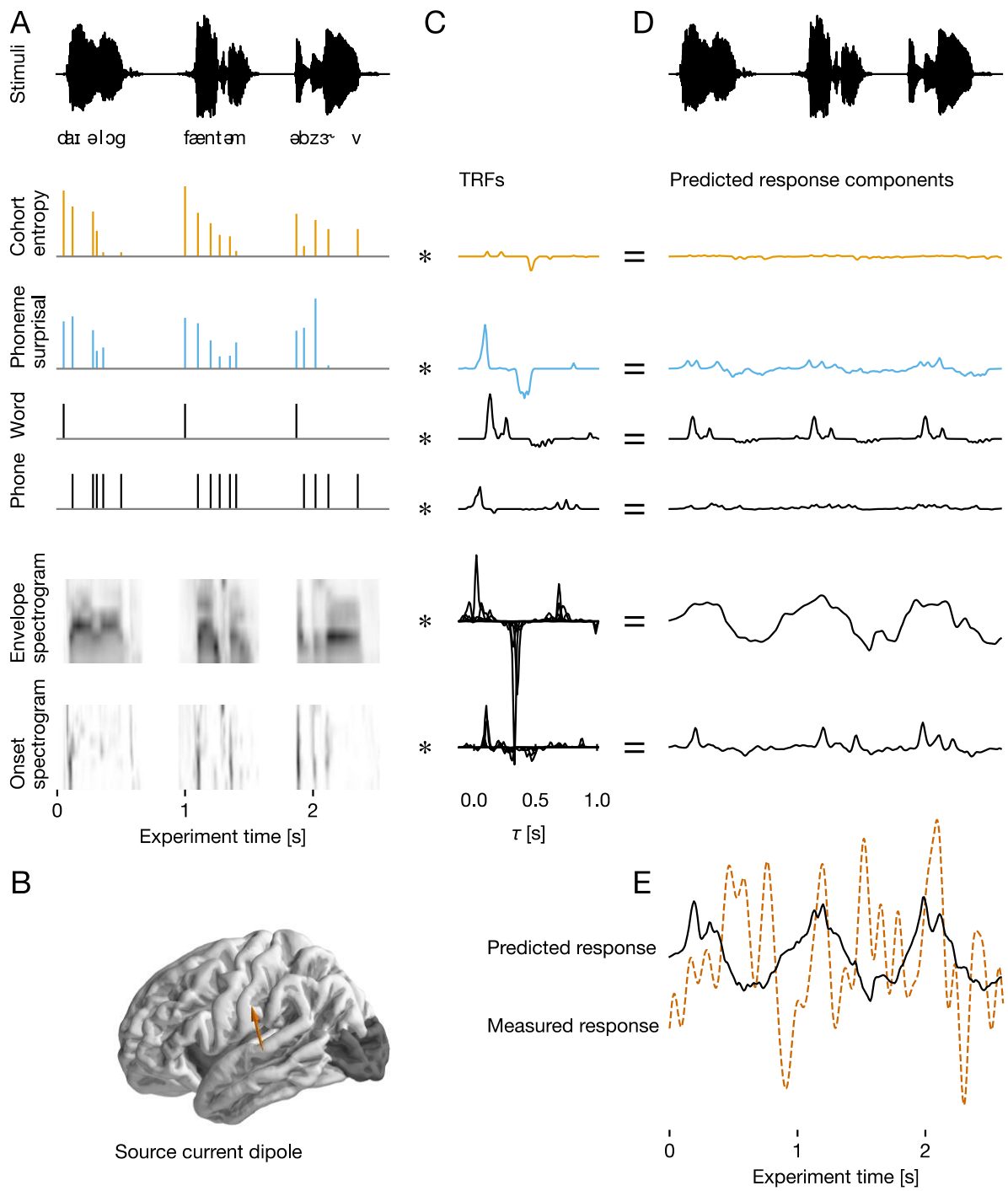

Figure 2. *Temporal response function analysis. Brain activity was modeled as continuous response to multiple variables describing the sequence of words.* ***(A)*** *Predictor variables used to describe the stimuli were all represented as time series. Cohort entropy and phoneme surprisal were modeled as impulses at phoneme onset, scaled by the relevant quantity. Covariates included word and phoneme onsets, an 8-band auditory spectrogram, and an 8-band auditory onset spectrogram.* ***(B)*** *Neural activity was quantified as distributed minimum norm current estimates, i.e., estimated current at a grid of dipoles covering the cortical surface. The analysis was restricted to the temporal, frontal, and parietal lobes (the dark shading indicates regions*

excluded from the analysis). One dipole from one representative subject is used in this figure for illustration. ***(C)*** *Temporal response functions (TRFs) were estimated using a coordinate descent algorithm to predict the neural signal from the predictor variables.* ***(D)*** *TRFs were estimated jointly, i.e., each TRF, convolved with its corresponding predictor variable time series, predicted a component of the neural activity. The sum of these component responses is the predicted brain response* ***(E)****. Model performance was evaluated by the proportion of the variability in the measured response that was explained by the predicted response.*

Participants heard a list of 1000 monomorphemic words with an inter-stimulus interval of 267 ms, and responded to randomly occurring semantic relatedness probes. Models were fit using 5-fold cross-validation in each subject separately. We evaluated the models by the proportion of variability they explained in the source-localized MEG recordings, correcting for multiple comparisons using threshold-free cluster enhancement (Smith & Nichols, 2009). Unless noted specifically, analyses were performed on the surface of the temporal, frontal and parietal lobes combined (see shaded area in **Figure 2B**).

Materials & Methods

Participants

We collected MEG data from 24 people. Sample size was chosen in accordance with the previous studies cited in Table 1. All participants were right-handed, native speakers of English, and seven were also native speakers of additional languages. None reported a history of neurological or linguistic impairment, brain injury, or hearing loss. All reported normal or corrected-to-normal vision. The procedure was approved by the University of Maryland Institutional Review Board and all participants provided written informed consent. Participants were compensated with their choice of $15 or 1 course credit per hour of participation. The full session (including another, unrelated study) lasted 2 hours.

One dataset was excluded before data processing because of participant fatigue and an earbud falling out during the experiment. After this exclusion, we computed accuracy on the semantic relatedness task and excluded any participant with accuracy lower than a cutoff one standard deviation below the mean. This excluded three of 23 participants. After preprocessing, two additional datasets were excluded due to excessive magnetic noise. 18 datasets are therefore included in our analysis.

Stimuli

Our stimuli were word recordings from the Massive Auditory Lexical Decision (MALD) database (Tucker et al., 2019), which includes the timing of phoneme boundaries from a forced aligner. The set of 1000 words we selected had no missing variables in the database and were monomorphemic per MALD, CELEX (Baayen et al., 1995), and first author judgment. We excluded all items with the following labels in MALD: Preposition, Interjection, Name, Unclassified, Conjunction, Pronoun, Determiner, Letter, Not, Ex, Article, To. We also removed items with the 10% lowest frequency values, and excluded homophones, inappropriate and particularly evocative words, and any item for which the pronunciation in the recording was noticeably divergent from American English. The full lists of stimuli and semantic relatedness probes (see below), as well as associated stimulus variables from MALD, are available on OSF (https://osf.io/u56ea/).

Procedure

The study was always the second of two experiments in a session. Before the MEG recording, we used a Polhemus 3SPACE FASTRAK to digitize participant head shapes as well as the positions of five affixed marker coils. These marker coils were used to record head position relative to the MEG sensors before and after each study in the session. We recorded continuous MEG data, inside a magnetically shielded room, with a 160-channel axial gradiometer whole-head system (Kanazawa Institute of Technology, Kanazawa, Japan). Our sampling rate was 1000 Hz, and we used an online 60 Hz notch filter and 200 Hz low-pass filter.

Participants lay supine and looked at a screen overhead, while holding a button box in each hand. They wore foam earbuds and volume was adjusted to their comfort level. We instructed participants that they would hear a long series of random words, and that they should simply listen to the words while watching for probe words that would randomly appear on the screen with a question mark. They were instructed to press a button (with left hand for No and right hand for Yes) to indicate whether the word on the screen was related in any way to the word they had heard just before it.

We used Presentation (Neurobehavioral Systems, Inc., www.neurobs.com) to present the experiment. Our parameter and scenario files are available on OSF (https://osf.io/u56ea/). There were 1000 auditory trials interspersed pseudo-randomly with 97 semantic relatedness probe trials. The amount of time between trials was 267 ms. A visual fixation cross was on screen continuously during auditory trials and during the inter-trial interval. Each auditory trial simply consisted of presentation of the auditory stimulus and lasted the length of the auditory stimulus. Visual probe trials were pseudo-randomly distributed with a maximum interlude of 20 trials between probes. The probe (e.g., “podium?”) stayed on the screen until the participant pressed a button to answer.

We selected this task so that it would apply equally well to all types of words, and because we did not want button presses to occur on critical trials (as would happen in, e.g., lexical decision). The probe trials for which we expected participants to answer “No” were selected randomly from the list of eligible words that we did not end up using for auditory trials. Probe trials for which we expected participants to answer “Yes” were synonyms taken from the WordNet (https://wordnet.princeton.edu) page of the preceding auditory item and were also monomorphemic so as not to be trivially distinguishable from “No” trials. There was no overlap between probe words and words used in auditory trials. Which auditory trials would be followed with a probe were randomly selected. “Yes” and “No” probes were equally distributed.

The experiment lasted roughly 17 minutes. There was no built-in break, but participants were instructed that if they wished to take a break, they should simply delay their button press on a probe trial.

Data preprocessing

We processed the data using mne-python version 0.22 (Gramfort et al., 2013, 2014) and Eelbrain 0.34 (Brodbeck et al., 2019).

During file conversion with mne-python's kit2fiff GUI, we excluded any faulty marker measurements. We co-registered each digitized head shape with the Freesurfer (Fischl, 2012) "fsaverage" brain, using mne-python's co-registration GUI. We first used rotation and translation to align the digitized head shape and average MRI by the three fiducial points. We then used rotation, translation, and 3-axis scaling to minimize the distance between digitized head shape and average MRI points using the iterative closest point (ICP) algorithm. Convergence was always achieved within 40 iterations. For one participant, outlying points on the digitized head shape were removed between fitting to the fiducials and applying ICP.

Flat channels were automatically removed, and we used temporal signal space separation (Taulu & Simola, 2006) for removal of extraneous artifacts, with a buffer duration of 10 seconds. We then band-pass filtered the recordings between 1 – 40 Hz (mne-python default settings) and used ICA (independent components analysis, with extended infomax method) for removal of ocular, cardiac, and other extraneous artifacts. Components were selected manually based on their topography and time-course. After removing artifactual ICA components, we further low-pass filtered the data at 20 Hz, cropped it from 1 s before the first word to 2 s after the last word, and down-sampled it to 100 Hz.

To compute a noise covariance matrix, we used two minutes of empty room data recorded before or after each session. We defined the source space on the white matter surface with a four-fold icosahedral subdivision, with 2562 sources per hemisphere. Orientation of the source dipoles was fixed perpendicular to the white matter surface. Continuous data were source localized with the regularized minimum norm estimator ($\lambda = 1/6$).

Analysis

Behavioral data

Mean accuracy was computed after the exclusion of one participant a priori. The mean number of correct probe responses was 73.6 (out of 97) with a standard deviation of 18.4. The number of correct probe responses was lower than one standard deviation below the mean for three participants, so they were excluded from further analysis. One participant answered 13 of 97 probes correctly. We kept this participant in the dataset because this was so far below chance that the only plausible explanation seemed to be that they had reversed which hand they were supposed to use to make Yes and No responses.

Predictors for neural data

For each stimulus variable of interest, a time series was created indicating the value of the predictor at each time point in the stimulus and aligned to the MEG data. For the acoustic predictors, the value of the predictor can vary continuously at each time point. For linguistic predictors, values are non-zero only at time points labeled as phoneme onsets, as determined by the forced alignments (i.e., linguistic predictors consist of impulses at phoneme onsets). Of these lexical predictors, the phoneme onset and word onset predictors each consist of binary impulses, while the predictors for cohort entropy and phoneme surprisal consist of impulses that are scaled continuously according to the variable value at that phoneme. Our study did not actually present a continuous stimulus (rather, individual words with short intervening pauses), but a single time series reflecting predictor values (or pauses) throughout the entire experiment could still be

created. Probe trials were modeled simply as silence. The timing of phoneme onsets was taken from the forced aligner information made available with the MALD recordings.

Acoustic spectrogram: a gammatone spectrogram (Heeris, 2018) was computed for each stimulus waveform with 256 channels regularly spaced in ERB space between 20 and 5000 Hz. These spectrograms were resampled to 100 Hz to match the MEG data and binned into eight equally spaced frequency bands.

Acoustic onset spectrograms: The high resolution gammatone spectrograms were processed with an algorithm for acoustic edge extraction (Brodbeck et al., 2020; Fishbach et al., 2001). The onset spectrograms were also resampled to 100 Hz and binned into eight bands.

Word onsets: Word onsets were represented as a single, equally valued impulse at the onset of every word, as determined from the forced alignments.

Phoneme onsets: Phoneme onsets, including only phonemes that were not also word onsets, were represented as equally valued impulses on a single predictor time series.

Phoneme surprisal and cohort entropy: these variables were calculated based on an implementation of the cohort model of word perception (Marslen-Wilson, 1987), as in Brodbeck, Hong, and Simon (2018). Initially, a dictionary was created combining frequency information from SUBTLEX (Brysbaert & New, 2009) with pronunciations (phoneme sequences) from the CMU pronouncing dictionary (Weide, 1994), adding any pronunciations from the stimuli that were missing from the CMU dictionary. This dictionary was then used to compute the set of words compatible with the input so far for each word at each phoneme position. These cohorts, together with the SUBTLEX frequencies, were used to compute a probability distribution over possible words for each phoneme position. The *cohort entropy* predictor contained an impulse at each phoneme onset, scaled by the entropy of that cohort. The *phoneme surprisal* predictor contained an impulse at each phoneme onset scaled by the surprisal of that phoneme, based on the posterior probability of that phoneme given the preceding phoneme's cohort.

TRF analysis

A multivariate temporal response function (mTRF) maps a set of predictor variables to a single outcome time series. Here, independent mTRFs were estimated for each subject and for each virtual current source of source-localized MEG data (see **Figure 2**). The neural response at time t, y_t is predicted jointly from N predictor time series, represented as $x_{i,t}$, convolved with a corresponding mTRF $h_{i,\tau}$ of length T:

$$\hat{y}_t = \sum_{i}^{N} \sum_{\tau}^{T} h_{i,\tau} \cdot x_{i,t-\tau}$$

mTRFs were generated from a basis of 50 ms wide Hamming windows centered at T=[-100,…,1000) ms. All responses and predictors were standardized by centering and dividing by the mean absolute value.

For a given set of predictors, the predictive power was estimated through 5-fold cross-validation. The continuous data and corresponding predictors were split into five contiguous partitions of

even length. The neural responses of each partition were predicted with an mTRF trained on the remaining four partitions to minimize $\ell 1$ error. Within each set of four training partitions, each partition in turn served as validation data once as four mTRFs were estimated based on coordinate descent with early stopping based on the validation data (David et al., 2007). The validation data was used to selectively stop training predictors when they caused an increase in error in the validation set. Those four mTRFs were then averaged to predict the responses to the unseen (fifth) test segment.

For evaluating the predictive power of the relevant predictors, phoneme surprisal and cohort entropy, we compared the predictive power of the full model with that of a model that was identical except for not including the predictor under investigation. Together with the cross-validation, this assures a conservative estimate of the unique predictive power of the predictor under investigation, while controlling for the predictive power of all the other variables. The anatomical maps of explanatory power of the two models were compared with a mass-univariate related measures *t*-test based on threshold-free cluster enhancement (TFCE) (Smith & Nichols, 2009), with a null distribution based on 10,000 random permutations of condition (model) labels.

For analysis of the TRFs, the five estimates of the TRFs from the five different test partitions were averaged in each subject. In order to visualize the TRF current over time, the TRF was restricted to the anatomical area in which the surprisal predictor significantly improved predictions ($p \leq .05$ based on TFCE). Within this area, and separately for each hemisphere and each participant, principal component analysis was applied to the virtual current dipoles, and only the first principal component was analyzed, i.e., a single spatial topography and corresponding time course for each participant. The advantage of this approach over others, such as root mean squared activity, is that the signed current direction can be visualized. Because the sign of a principal components is arbitrary, the components were aligned across subjects such that the average current vector was pointing upward. For components whose average current vector pointed downward, both component and time-course were multiplied by -1.

TRF time-course was then evaluated in each hemisphere using a mass-univariate one-sample *t*-test with TFCE, with the null hypothesis that the average current direction is random (i.e., not different from 0). The null distribution was based on the maximum statistic in 10,000 random permutations of the signs. To test for hemispheric differences, a mass-univariate repeated measures *t*-test with the same parameters was used.

Comparison with connected speech

For this comparison, data from 12 participants listening to 47 minutes of a non-fiction audiobook were used (for more details see Brodbeck et al. (2021)). Data were acquired on the same MEG equipment and with analogous procedures, with one exception: For estimation of the mTRF models, data were split into four partitions instead of five. This was done to speed up computations (requiring training of fewer models) and because the longer recording resulted in more training data per participant. Audiobook stimuli were labeled using the Montreal Forced Aligner (McAuliffe et al., 2017), and predictor variables were generated as for the single-word data.

Results

To ensure that responses reflect attentive lexical processing, we applied behavioral exclusion criteria (see Methods). Subjects included in the analyses presented here responded accurately to at least 69% of relatedness probes (group mean 82.9%).

Our first question was whether first phonemes should be excluded from the phoneme surprisal and cohort entropy estimates, as suggested by a lack of first phoneme cohort effects reported by Brodbeck, Hong, and Simon (2018) and Gaston and Marantz (2018). To answer this question, we compared the model treating all phonemes uniformly (**Figure 2**) to a model in which surprisal and entropy at the first phoneme are modeled as separate predictors from surprisal and entropy at non-initial phonemes. The more complex model, in which they are modeled separately, was not significantly better (t_{max} = 2.74, p = .341, multiple comparison correction in temporal lobes only). We therefore proceeded with the simpler model in which initial phonemes are not modeled separately (as shown in **Figure 2**). Acoustic and segmentation variables were always controlled for (see Methods).

We then tested our primary question: do phoneme surprisal and cohort entropy improve the estimated neural response in a single-word design? We found that indeed, a model with phoneme surprisal was significantly better than a model without it (p < 0.001). However, comparison with a model lacking cohort entropy led to no significant difference (p = .260, see **Figure 3**). The model improvement due to surprisal (i.e., the explanatory power of surprisal) was significantly larger than that due to entropy (p = .007).

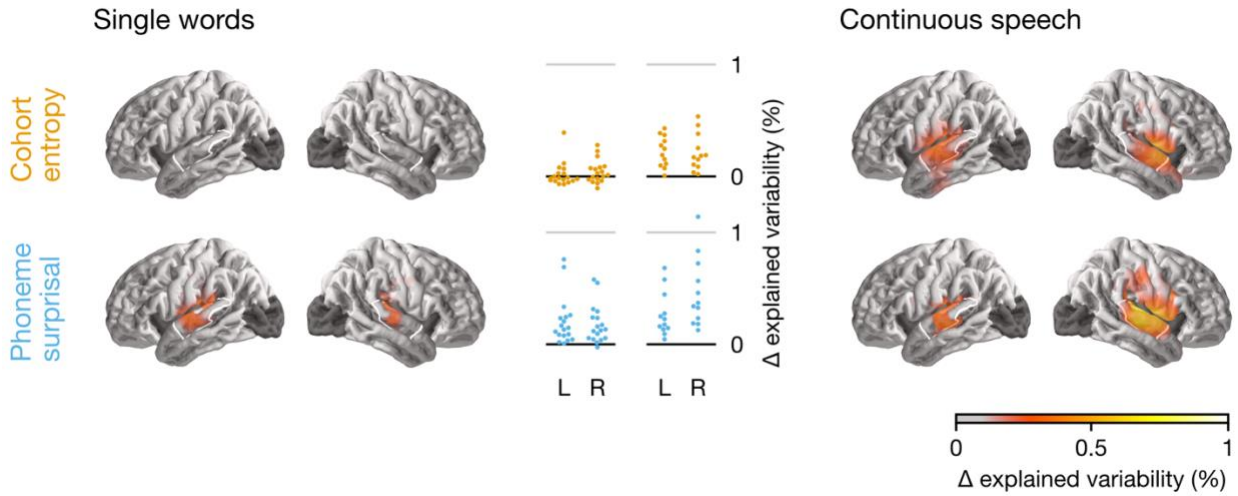

Figure 3. *Model evaluation and comparison to continuous speech. The anatomical plots at left and right show regions where the given predictor significantly improved the model fit. The white outline indicates an anatomical region of interest (ROI) defined as the posterior 2/3 of the superior temporal gyrus. The swarm plots (middle) show average proportion of variability in that ROI that is uniquely explained by entropy or surprisal, respectively. Each dot represents one participant. While surprisal improves the model fit in both experiments in almost all participants, entropy does so only in the continuous-speech data. Explained variability (explanatory power) is expressed as percentage of the maximum variability explained by the full model in the single-word data.*

This finding contrasts with previously reported results in continuous speech (see Table 1). To address this apparent difference, we compared our single-word data to an existing continuous-speech dataset acquired with the same MEG scanner (Brodbeck et al., 2021), consisting of recordings from 12 participants listening to 47 minutes of an audiobook, using closely matched analysis methods (**Figure 3**). For the continuous-speech data, phoneme surprisal significantly improved the model ($p < .001$) and cohort entropy did as well ($p < .001$). In the whole brain analysis, the explanatory power of phoneme surprisal and cohort entropy did not differ significantly ($p = .720$).

To confirm this difference between experiments statistically, we extracted the mean of the model fit metric in a region of interest (ROI) defined as the posterior two thirds of the superior temporal gyrus of each hemisphere. This value did not differ between the left and right hemisphere ROIs in any of the four categories (surprisal/entropy, single words/continuous speech; all $t \leq 1.74$, $p \geq .110$), so we averaged the values for the two hemispheres. We then calculated the ratio between the predictive power of entropy and surprisal and compared this ratio for continuous speech and single words. Such a test is unlikely to be affected by differences in the size of the datasets. This ratio was significantly higher in continuous speech than for single words (continuous speech $M = 0.68$, $SD = 0.45$; single words $M = 0.10$, $SD = 0.59$; $t_{28} = 2.80$, $p = .009$). Consistent with this, effect sizes for predictive power in the ROI were large for surprisal in both paradigms (single words: $d = 1.62$; connected speech: $d = 2.14$) but for entropy only in connected speech ($d = 1.72$) and not in single words ($d = 0.39$).

Finally, we examined the nature of the estimated response functions for phoneme surprisal in the single-word dataset **(Figure 4)**. The TRF analysis was restricted to a mirror-symmetric anatomical region, based on the area in which surprisal significantly improved the model fit in at least one hemisphere. Because the TRF was relatively well captured by a single topography, we extracted only the first principal component of the TRF for each participant and each hemisphere (a parallel analysis using the response magnitude led to the same conclusions). **Figure 4A** shows the average of the first principal component for each subject. The result in both hemispheres was consistent with a current dipole in auditory cortex, indicated by the arrows in **Figure 4A**. The time-course in the two hemispheres **(Figure 4B)** was analyzed with mass-univariate *t*-tests, correcting for the time range from 0 to 1000 ms. In both hemispheres, an early peak around 90 ms was followed by more extended activity of opposite current direction, starting around 280 ms. Even though activity in the early peak did not reach significance in the right hemisphere, the difference between hemispheres, based on a mass-univariate related measures *t*-test, was not significant ($p = .063$, at 70 ms).

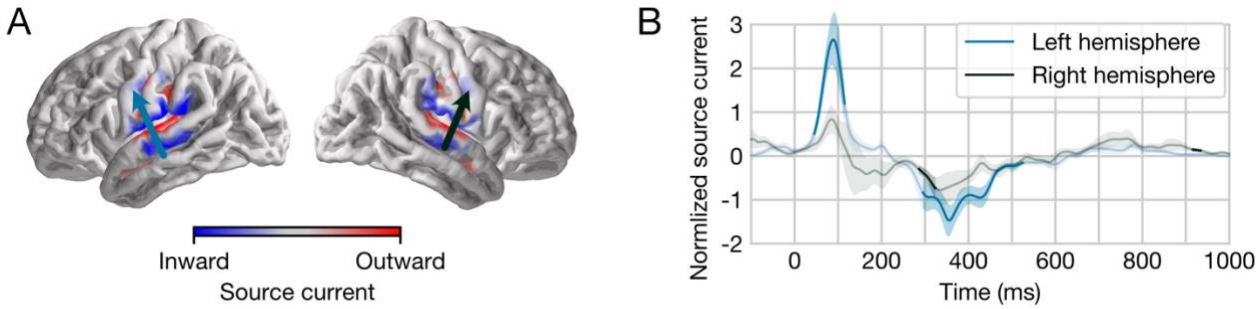

Figure 4. *TRF results for phoneme surprisal. The TRF is analyzed using the first principal component in each subject.* ***(A)*** *The average first principal component across subjects. The average current direction (indicated by arrows) is consistent with auditory cortex activity.* ***(B)*** *The time-course of the component separately for the left and the right hemisphere. Solid line segments indicate time ranges in which the respective TRF is significantly different from zero.*

Discussion

This study examined cohort entropy and phoneme surprisal effects in a single-word paradigm using a temporal response function analysis, modeling both acoustic and linguistic predictors of neural activity. We found that phoneme surprisal is a significant predictor of neural activity during speech recognition, as have many previous studies (Brodbeck et al., 2018, 2021; Donhauser & Baillet, 2020; Ettinger et al., 2014; Gagnepain et al., 2012; Gaston & Marantz, 2018; Gillis et al., 2021; Gwilliams et al., 2020; Gwilliams & Marantz, 2015). The spatial distribution of the effect along the superior temporal gyrus is also consistent with previous work. The TRF for phoneme surprisal in our study appears to peak twice, in line with Gwilliams and Marantz (2015), Gaston and Marantz (2018), and Brodbeck et al. (2021).

In contrast to the robust effect of phoneme surprisal, we did not observe a significant effect of cohort entropy. In a direct comparison to our single-word dataset, we analyzed a continuous-speech dataset (Brodbeck et al., 2021) in the same manner, and found effects of both phoneme surprisal and cohort entropy. The direct comparison of these two datasets substantiates our generalization about the existing literature, that cohort entropy effects are weak or non-existent in studies that use single words or short phrases, while they are robust in studies that use continuous, naturalistic speech as stimuli.

How could this dissociation between phoneme surprisal and cohort entropy occur? As reviewed in the Introduction, it is frequently assumed that speech input triggers relatively automatic and incremental activation of phoneme, wordform, lexical-syntactic, and conceptual units, but this would predict cohort entropy effects for any task involving word recognition. In the following sections, we hypothesize (1) that brain responses related to phoneme surprisal and cohort entropy arise from different levels of representation or different sub-processes and (2) that their dissociation therefore implies a break in the automatic sequence of processing involved in word recognition.

Non-automaticity in the lexical access sequence

The pattern of dissociation that we observed could have several different explanations, contingent on the precise neural processes indexed by cohort entropy and phoneme surprisal. In

Figure 5A, we reproduce our illustration in **Figure 1** of a fully automatic processing sequence in response to each phoneme of speech input. In **Figure 5B-D**, we illustrate alternative partial versions of this sequence that might better represent what occurs incrementally in single-word paradigms that do not elicit cohort entropy effects. It is possible that the decoupled processes do not occur at all in single-word processing; alternatively, they could be engaged much later or engaged in a less strictly incremental, time-locked manner rather than on a phoneme-by-phoneme basis.

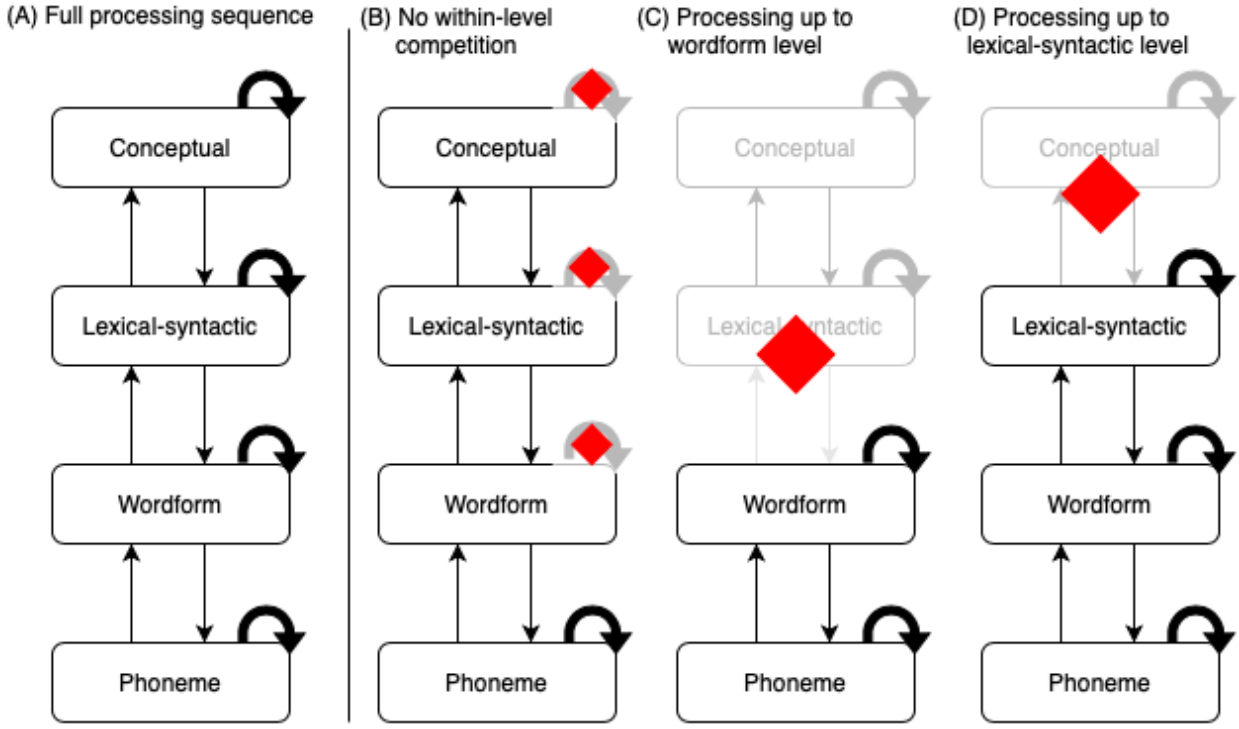

Figure 5. (A) *Fully automatic processing sequence in which both phoneme surprisal and cohort entropy effects arise.* ***(B)-(D)*** *Proposed partial processing sequences in which phoneme surprisal but not cohort entropy effects occur. Red diamonds indicate processes or levels of representation that might be delayed or suspended from incremental (phoneme-locked) processing during recognition of single words. As in* ***Figure 1****, straight arrows indicate connections between levels of representation. Curved arrows indicate a within-level competition/selection process.*

One possible explanation is based on the reasoning that cohort entropy is specifically a measure of the amount of lexical competition occurring (Gagnepain et al., 2012). We can imagine a scenario in which initial activation of multiple lexical candidates is automatic, but in which the competitive process of winnowing out the weaker ones is only applied when rapid selection of a single best candidate is particularly helpful or necessary for the task at hand. Accordingly, phoneme surprisal effects might require only *activation* of, e.g., the wordform level of representation, rather than the competition process that occurs within that level (a scenario illustrated by **Figure 5B**, in which within-level competition processes are not occurring above the phoneme level). In contrast, cohort entropy effects would reflect the competitive selection process that allows a single best candidate to be identified as early as possible, and this process might only be engaged when faced with the time pressure of processing connected speech.

Another possibility is that phoneme surprisal and cohort entropy effects reflect different levels of representation which are not all automatically accessed to the same degree. Access to 'lower' levels of representation like phoneme or wordform representations might be more automatic, whereas access to 'higher' levels of representation like lexical-syntactic or conceptual units might be more dependent on context and task demands. For instance, surprisal effects might require only wordform-level activation, while cohort entropy effects might require lexical-syntactic activation or higher. Similarly, phoneme surprisal effects could implicate up to lexical-syntactic activation while cohort entropy effects require conceptual activation or higher. These two possibilities are illustrated in **Figure 5C** and **Figure 5D**, respectively. Consistent with such an explanation, semantic priming from partial wordforms seems to be more reliable in connected speech (Zwitserlood, 1989) than in a single-word lexical-decision paradigm, where form-based priming predominates (Gaskell & Marslen-Wilson, 2002). Even within a single-word paradigm, Bentin et al. (1993) argue that the extent to which a task requires semantic processing can influence the degree of semantic priming that occurs, as indexed by the N400 response.

Though less likely, we can acknowledge two alternative explanations in which phoneme surprisal effects are not actually related to wordform representations analogous to the ones used to calculate phoneme surprisal. We consider these less likely because they would imply an absence of incremental wordform-level processing in the single-word tasks, despite behavioral evidence to the contrary. One possibility is that apparent phoneme surprisal effects arise due to prelexical phonotactic processing, involving representations sensitive to the probability of phoneme sequences in the language independent of wordform representations. The second possibility arises from the proposal of Norris & McQueen (2008) that 'off-line' perceptual learning could lead to wordform frequency effects on phoneme probability without concurrent wordform activation causing online top-down effects. In either of these scenarios, a phoneme surprisal effect does not necessarily imply wordform activation; cohort entropy could reflect anything at the wordform level or above. Similarly, it remains possible that correlations between neural activity and cohort entropy are not driven by lexical competition or uncertainty per se but by a secondary process that is sensitive to lexical competition or uncertainty. If that process is only engaged by continuous speech, cohort entropy effects would also appear to be modulated.

Single words vs. continuous speech

What are the differences between single-word paradigms and continuous speech that would make any of the distinctions described in the previous section possible? First, the reliable presence of pauses between words in single-word studies may constitute a key change in task demands, by leaving sufficient time for full lexical access to occur after wordform offset and before the next wordform begins and thus reducing the necessity for incremental processing. Early competitive selection might be unnecessary, and/or access to higher-level syntactic and conceptual units could be deferred until the pause makes the auditory wordform uniquely identifiable. Among the single-word studies we have reviewed, the pause detection (Gagnepain et al., 2012) and nonword detection (Kocagoncu et al., 2017) tasks incorporate lengthy inter-stimulus intervals averaging 2000 ms, and the lexical decision studies (Brennan et al., 2014; Ettinger et al., 2014; Gwilliams & Marantz, 2015; Lewis & Poeppel, 2014) wait for a participant response after each word. Our study used a shorter but still considerable inter-stimulus interval of 267 ms with only occasional semantic relatedness probes and also did not find a cohort entropy effect.

Second, the syntactic and semantic structure in continuous speech provides another motivation for incremental processing: rapid access to lexical and conceptual content for the current word provides information that might aid recognition of the subsequent word. This rationale for rapid processing is absent in single-word paradigms that lack structure. Even beyond not requiring speed in lexical or conceptual access, the tasks employed in single-word paradigms may in some cases not require lexical or conceptual access at all. For instance, our task involved semantic relatedness judgements with written probes. It is conceivable that this task might be solved successfully by temporarily 'buffering' the input from each word as a form-based representation, and only accessing conceptual representations if a probe occurs. By contrast, the speed of continuous speech, its many between-word dependencies, and the imperative to build sentence-level and message-level interpretations could be what drive competition or incremental higher-level activation (and therefore cohort entropy effects) in naturalistic paradigms.

We might expect that cohort entropy effects could be observed for single words if a task were designed such that earlier identification of the word is encouraged and incremental higher-level activation becomes more advantageous, whether via the elimination of pauses or the addition of some higher-level structure. Likewise, pauses could be added to continuous speech. The three-word phrases used by Gaston and Marantz (2018) (e.g., "to chew gum," "the shredder broke") are an interesting test of these hypotheses, as they lack within-phrase pauses and have syntactic and semantic structure. Nevertheless, Gaston and Marantz did not find cohort entropy effects when their cohort entropy variables were evaluated in the same model as phoneme surprisal. This suggests that only longer sequences of continuous speech elicit cohort entropy effects, and therefore that a buffering process may play a mediating role here.

Another line of investigation for understanding what drives neural cohort effects might involve the contrast between monomorphemic and multimorphemic words. The two types of stimuli can be closely matched in length, but only multimorphemic words can be viewed as structured sequences of units of meaning of the kind that might encourage more incremental processing. The inclusion of multimorphemic words in a single-word study could thus motivate earlier selection and higher-level activation so that initial morphemes can be recognized in time to begin processing any potential subsequent morphemes. In Table 1, we noted that all single-word studies that do not include multimorphemic words also do not report cohort entropy effects (Brennan et al., 2014; Gagnepain et al., 2012; Lewis & Poeppel, 2014). This is true of our study as well. Among the single-word studies that do report cohort entropy effects, albeit without controlling for phoneme surprisal, Ettinger et al. (2014) include multimorphemic words and Kocagoncu et al. (2017) do not indicate whether multimorphemic words are included in their stimuli or not. This factor deserves further investigation. In particular, it is unclear whether hypothesized cohorts should be constructed on a morpheme-by-morpheme or wordform-by-wordform basis.

Implications

If auditory word recognition in most single-word studies proceeds in the manner we have proposed, with candidate selection or higher-than-wordform-level processing delayed or suspended entirely, there are two major implications. The first is that the cascading, incremental access process is not automatic but rather is motivated by time pressure and modulable with the extent of that time pressure. The second is that auditory word recognition in many single-word

studies may differ fundamentally from the process most researchers assume they are studying (that is, speech recognition in natural contexts). This would invite re-interpretation of existing neural and behavioral data and would motivate increased use of more naturalistic designs in future work, or identification of changes to single-word paradigms that would drive cohort entropy effects so that these paradigms can be used with more confidence that they are representative of the processing of natural, connected speech.

Conclusion

Our goal in this study was to evaluate whether an assumption of parallelism is warranted between recognition of single words and word recognition in natural connected speech. We also intended to establish a better understanding of what drives phoneme surprisal and cohort entropy effects, while modeling the speech stimulus as thoroughly as current methods allow. We directly compared single-word and continuous-speech data from MEG and demonstrated the occurrence of phoneme surprisal effects in both paradigms but a cohort entropy effect only in continuous speech, consistent with patterns in the existing literature. We proposed that this is because phoneme surprisal effects arise from the activation of a lower level of representation while cohort entropy effects arise from a competition process or higher level of representation whose engagement is delayed or does not occur in single-word paradigms. This dissociation suggests that the sequence of processing triggered by speech input is not automatic and the extent to which competition processes or higher levels of representation are engaged depends on the nature of the stimulus or experimental task. This study has also helped validate the TRF approach as a promising method for future work in single-word paradigms.

References

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227. https://doi.org/10.1016/j.conb.2017.08.010

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, *38*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX2 LDC96L14 [Web Download]*. Linguistic Data Consortium.

Baayen, H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon*, *2*(3), 419–463. https://doi.org/10.1075/ml.2.3.06baa

Balling, L. W., & Baayen, H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*(1), 80–106. https://doi.org/10.1016/j.cognition.2012.06.003

Bentin, S., Kutas, M., & Hillyard, S. A. (1993). Electrophysiological evidence for task effects on semantic priming in auditory word processing. *Psychophysiology*, *30*(2), 161–169. https://doi.org/10.1111/j.1469-8986.1993.tb01729.x

Bien, H., Baayen, R. H., & Levelt, W. J. M. (2011). Frequency effects in the production of Dutch deverbal adjectives and inflected verbs. *Language and Cognitive Processes*, *26*(4–6), 683–715. https://doi.org/10.1080/01690965.2010.511475

Brennan, J., Lignos, C., Embick, D., & Roberts, T. P. L. (2014). Spectro-temporal correlates of lexical access during auditory lexical decision. *Brain and Language*, *133*, 39–46. https://doi.org/10.1016/j.bandl.2014.03.006

Brodbeck, C., Bhattasali, S., Cruz Heredia, A., Resnik, P., Simon, J. Z., & Lau, E. (2021). *Parallel processing in speech perception: Local and global representations of linguistic context*. https://doi.org/10.1101/2021.07.03.450698

Brodbeck, C., Das, P., Brooks, T., & Reddigari, S. (2019). *Eelbrain 0.31* (v0.31) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3564850

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*, *28*(24), 3976-3983.e5. https://doi.org/10.1016/j.cub.2018.10.042

Brodbeck, C., Jiao, A., Hong, L. E., & Simon, J. Z. (2020). Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. *PLOS Biology*, *18*(10), e3000883. https://doi.org/10.1371/journal.pbio.3000883

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word Familiarity and Frequency in Visual and Auditory Word Recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(6), 1084–1096. https://doi.org/10.1037/0278-7393.16.6.1084

Dahan, D., & Magnuson, J. S. (2006). Spoken Word Recognition. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (2nd ed., pp. 249–283). Elsevier. https://doi.org/10.1016/B978-012369374-7/50009-2

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time Course of Frequency Effects in Spoken-Word Recognition: Evidence from Eye Movements. *Cognitive Psychology*, *42*(4), 317–367. https://doi.org/10.1006/cogp.2001.0750

David, S. V., Mesgarani, N., & Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network: Computation in Neural Systems*, *18*(3), 191–212. https://doi.org/10.1080/09548980701609235

Di Liberto, G. M., Wong, D., Melnik, G. A., & de Cheveigné, A. (2019). Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *NeuroImage*, *196*, 237–247. https://doi.org/10.1016/j.neuroimage.2019.04.037

Donhauser, P. W., & Baillet, S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*, *105*(2), 385-393.e9. https://doi.org/10.1016/j.neuron.2019.10.019

Ettinger, A., Linzen, T., & Marantz, A. (2014). The role of morphology in phoneme prediction: Evidence from MEG. *Brain and Language*, *129*, 14–23. https://doi.org/10.1016/j.bandl.2013.11.004

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Fishbach, A., Nelken, I., & Yeshurun, Y. (2001). Auditory Edge Detection: A Neural Model for Physiological and Psychoacoustical Responses to Amplitude Transients. *Journal of Neurophysiology*, *85*(6), 2303–2323. https://doi.org/10.1152/jn.2001.85.6.2303

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal Predictive Codes for Spoken Words in Auditory Cortex. *Current Biology*, *22*(7), 615–621. https://doi.org/10.1016/j.cub.2012.02.015

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology*, *45*(2), 220–266. https://doi.org/10.1016/S0010-0285(02)00003-8

Gaston, P., & Marantz, A. (2018). The time course of contextual cohort effects in auditory processing of category-ambiguous words: MEG evidence for a single "clash" as noun or verb. *Language, Cognition and Neuroscience*, *33*(4), 402–423. https://doi.org/10.1080/23273798.2017.1395466

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., & Brodbeck, C. (2021). *Neural markers of speech comprehension: Measuring EEG tracking of linguistic speech representations, controlling the speech acoustics* [Preprint]. Neuroscience. https://doi.org/10.1101/2021.03.24.436758

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*(Supplement C), 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*(4), 267–283. https://doi.org/10.3758/BF03204386

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2020). *Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.04.04.025684

Gwilliams, L., & Marantz, A. (2015). Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words. *Brain and Language*, *147*, 1–13. https://doi.org/10.1016/j.bandl.2015.04.006

Heeris, J. (2018). *Gammatone Filterbank Toolkit* (0626328ef7c31d3b33214db2fdcd52e8601eb4c5) [Computer software]. https://github.com/detly/gammatone

Kemps, R. J. J. K., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, H. (2005). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes*, *20*(1–2), 43–73. https://doi.org/10.1080/01690960444000223

Kocagoncu, E., Clarke, A., Devereux, B. J., & Tyler, L. K. (2017). Decoding the Cortical Dynamics of Sound-Meaning Mapping. *The Journal of Neuroscience*, *37*(5), 1312–1319. https://doi.org/10.1523/JNEUROSCI.2858-16.2016

Lewis, G., & Poeppel, D. (2014). The role of visual representations during the lexical access of spoken words. *Brain and Language*, *134*, 1–10. https://doi.org/10.1016/j.bandl.2014.03.008

Magnuson, J. S. (2016). Mapping spoken words to meaning. In M. G. Gaskell & J. Mirkovic (Eds.), *Speech Perception and Spoken Word Recognition* (pp. 76–96). Routledge.

Magnuson, J. S., Mirman, D., & Myers, E. (2013). Spoken Word Recognition. In D. Reisberg (Ed.), *Oxford Handbook of Cognitive Psychology* (pp. 412–441). Oxford University Press.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*, 1–71.

McAllister, J. M. (1988). The use of context in auditory word recognition. *Perception & Psychophysics*, *44*(1), 94–97. https://doi.org/10.3758/BF03207482

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017, August). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of the 18th Conference of the International Speech Communication Association*.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McQueen, J. M. (2007). Eight questions about spoken word recognition. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 36–54). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198568971.013.0003

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234. https://doi.org/10.1016/0010-0277(94)90043-4

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Smith, S., & Nichols, T. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Taulu, S., & Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine and Biology*, *51*(7), 1759–1768. https://doi.org/10.1088/0031-9155/51/7/008

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, *51*(3), 1187–1204. https://doi.org/10.3758/s13428-018-1056-1

Weide, R. (1994). *CMU pronouncing dictionary*. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Wurm, L. H., Ernestus, M. T. C., Schreuder, R., & Baayen, H. (2006). Dynamics of the auditory comprehension of prefixed words: Cohort entropies and Conditional Root Uniqueness Points. *The Mental Lexicon*, *1*(1), 125–146. https://doi.org/10.1075/ml.1.1.08wur

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 1–14. https://doi.org/10.1037/0278-7393.32.1.1

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, *32*(1), 25–64. https://doi.org/10.1016/0010-0277(89)90013-9