Parallel processing in speech perception: Local and global representations of linguistic context

Christian Brodbeck[1,2,*], Shohini Bhattasali[3,4], Aura Cruz Heredia[3,5], Philip Resnik[3,4], Jonathan Z. Simon[2,6,7] & Ellen Lau[3]

1) Department of Psychological Sciences, University of Connecticut, Storrs, CT, U.S.A.
2) Institute for Systems Research, University of Maryland, College Park, Maryland, U.S.A.
3) Department of Linguistics, University of Maryland, College Park, Maryland, U.S.A.
4) Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland, U.S.A.
5) Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.
6) Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland, U.S.A.
7) Department of Biology, University of Maryland, College Park, Maryland, U.S.A.

* christianbrodbeck@me.com

Keywords: temporal response functions, entropy, surprisal, MEG

Abstract

Speech processing is highly incremental. It is widely accepted that listeners continuously use the linguistic context to anticipate upcoming concepts, words and phonemes. However, previous evidence supports two seemingly contradictory models of how predictive cues are integrated with bottom-up evidence: Classic psycholinguistic paradigms suggest a two-stage model, in which acoustic input is represented fleetingly in a local, context-free manner, but quickly integrated with contextual constraints. This contrasts with the view that the brain constructs a single unified interpretation of the input, which fully integrates available information across representational hierarchies and predictively modulates even earliest sensory representations. To distinguish these hypotheses, we tested magnetoencephalography responses to continuous narrative speech for signatures of unified and local predictive models. Results provide evidence for some aspects of both. Local context models, one based on sublexical phoneme sequences, and one based on the phonemes in the current word alone, do uniquely predict some part of early neural responses; at the same time, even early responses to phonemes also reflect a unified model that incorporates sentence level constraints to predict upcoming phonemes. Neural source localization places the anatomical origins of the different predictive models in non-identical parts of the superior temporal lobes bilaterally, although the more local models tend to be right-lateralized. These results suggest that speech processing recruits both local and unified predictive models in parallel, reconciling previous disparate findings. Parallel models might make the perceptual system more robust, facilitate processing of unexpected inputs, and serve a function in language acquisition.

Introduction

Acoustic events in continuous speech occur at a rapid pace, and listeners face pressure to process the speech signal rapidly and incrementally[1]. One strategy that listeners employ to achieve this is to organize internal representations in such a way as to minimize the processing cost of future language input[2]. This is reflected in a variety of measures that suggest that more predictable words are easier to process[3–5]. For instance, spoken words are recognized more quickly when they are heard in a meaningful context[6], and words that are made more likely by the context are associated with reduced neural responses, compared to less expected words[7–11]. This contextual facilitation occurs broadly and is sensitive to language statistics[12–14] as well as discourse level meaning[15,16].

Words are predictable because they occur in sequences that form meaningful messages. Similarly, phonemes are predictable because they occur in sequences that form words. For example, after hearing the beginning /ɹɪv/, /ɝ/ would be a likely continuation forming *river*; /i/ would be more surprising, because *riviera* is a less frequent word, whereas /ʊ/ would be highly surprising because there are no common English words starting with that sequence. Phonemes that are thus inconsistent with known word forms elicit a mismatch response[17], and responses to valid phonemes are proportionately larger the more surprising the phonemes are[18–20]. Predictive processing is not restricted to linguistic representations, as even responses to acoustic features in early auditory cortex reflect expectations based on the acoustic context[21,22].

Thus, there is little doubt that the brain uses context to facilitate processing of upcoming information, at multiple levels of representation. Here we investigate a fundamental question about the underlying cognitive organization: Does the brain develop a single, unified representation of the input? In other words, one representation that is consistent across hierarchical levels, effectively propagating information from the sentence context across hierarchical levels to anticipate even low-level features of the sensory input such as phonemes? Or do cognitive subsystems differ in the extent and kind of context they use to interpret their input? This question has appeared in different forms, for example in early debates about whether sensory systems are modular[23] or whether sensory input and contextual constraints are combined immediately in speech perception[6,24]. A similar distinction has also surfaced more recently between the local and global architectures of predictive coding[25].

A strong argument for a unified, globally consistent model comes from Bayesian frameworks, which suggest that, for optimal interpretation of imperfect sensory signals, listeners ought to use the maximum amount of information available to them to compute a prior expectation for upcoming sensory input[26,27]. An implication is that speech processing is truly incremental, with a unified linguistic representation that is updated at the phoneme (or an even lower) time scale[5]. Such a unified representation is consistent with empirical results suggesting that word recognition can bias subsequent phonetic representations[28], that listeners weight cues like a Bayes-optimal observer during speech perception[29,30], and that they immediately interpret incoming speech with regard to communicative goals[31,32]. A recent implementation proposed for such a model is the global variant of hierarchical predictive coding, which assumes a cascade of generative models predicting sensory input from higher level expectations[25,33,34]. However, a unified model is also assumed by classical interactive models of speech processing, which rely on cross-hierarchy interactions to generate a globally consistent interpretation of the input[35–37].

However, there is also evidence for incomplete use of context in speech perception. Results from cross-modal semantic priming suggest that, during perception of a word, initially multiple meanings are activated regardless of whether they are consistent with the sentence context or not, and contextually appropriate meanings only come to dominate at a later stage[38,39]. Similarly, eye tracking suggests that lexical processing activates candidates that should be excluded by the syntactic context[40]. Such findings can be interpreted as evidence for a two-stage model, in which an earlier retrieval process operates without taking into account the wider sentence context, and only a secondary process of selection determines the best fit with context[41]. Similarly, experiments with non-words suggest that phoneme sequence probabilities can have effects separate from lexical processing[42,43]. However, it is also possible that such effects occur only due to the unnaturalness of experimental tasks. For example, in the cross-modal priming task, listeners might come to expect a visual target which is not subject to sentence context constraints, and thus change their reliance on that context.

Finally, a third possibility is that a unified model coexists with more local models of context, and that they operate in a parallel fashion. For example, it has been suggested that the two hemispheres differ with respect to their use of context, with the left hemisphere relying heavily on top-down predictions, and the right hemisphere processing language in a more bottom-up manner[44].

Distinguishing among these possibilities requires a task that encourages naturalistic engagement with the context, and a non-intrusive measure of linguistic processing. To achieve this, we analyzed magnetoencephalography (MEG) responses to continuous narrative speech. Previous work using a similar paradigm has tested either only for a local or only for a unified context model, by either using only the current word up to the current phoneme as context[45,46] or by using predictions from a complete history of phonemes and words[47]. However, because these two context models include overlapping sets of constraints, their predictions are correlated and they need to be assessed jointly. Furthermore, some architectures predict that both kinds of context model should affect brain responses separately. For example, a two-stage architecture predicts an earlier stage of lexical processing that is sensitive to lexical statistics only, and a later stage that is sensitive to the global sentence context. Here we directly test such possibilities by comparing the ability of different context models to jointly predict brain responses.

Expressing the use of context through information theory

The sensitivity of speech processing to different definitions of context is formalized through conditional probability distributions (Figure 1). Each distribution reflects an interpretation of ongoing speech input, at a given level of representation. We here use word forms and phonemes as units of representation (Figure 1-A), but this is a matter of methodological convenience, and similar models could be formulated using a different granularity[5]. Figure 1-B shows an architecture in which each level uses local information from that level, but information from higher levels does not affect beliefs at lower levels. In this architecture, phonemes are classified at the sublexical level based on the acoustic input and possibly a local phoneme history. The word level decodes the current word from the incoming phonemes, but without access to the multi-word context. Finally, the sentence level updates the sentence representation from the incoming word candidates, and thus selects those candidates that are consistent with the sentence context. In such a model, apparent top-down effects such as perceptual restoration of noisy input[48,49] are generated at higher level decision stages rather than at the initial perceptual representations[50]. In contrast, Figure 1-C illustrates the hypothesis of a unified or global context model, in which priors at lower levels take advantage of information available at the higher levels. Here, the sentence context is used in decoding the current word by directly altering the prior over word candidates, and this sentence-appropriate prior is in turn used to alter expectations for upcoming phonemes.

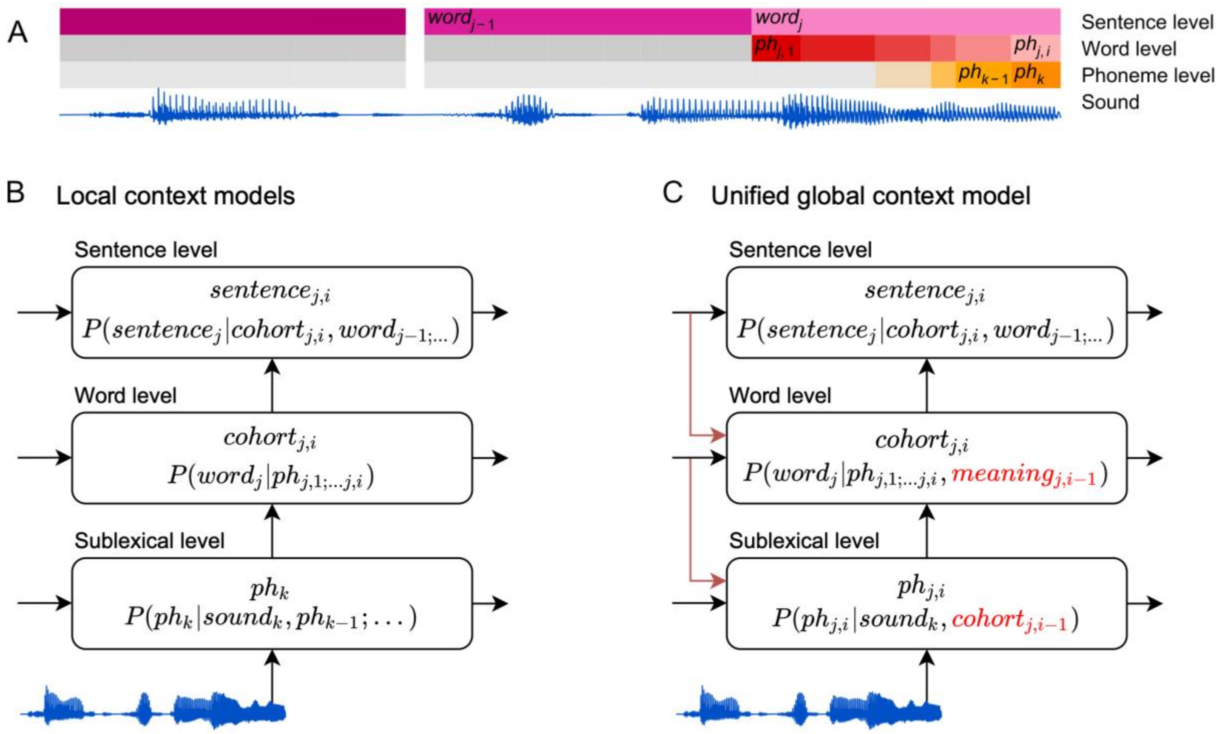

Figure 1. **Information flow in local and unified architectures for speech processing**
(A) Schematic characterization of the linguistic units used to characterize speech. The same phoneme can be invoked as part of a sublexical phoneme sequence, ph_k, or as part of $word_j$, $ph_{j,i}$.
(B) Each box stands for a level of representation, characterized by its output and a probability distribution describing the level's use of context. For example, the sublexical level's output is an estimate of the current phoneme, ph_k, and the distribution for ph_k is estimated as probability for different phonemes based on the sound input and a sublexical phoneme history. At the sentence level, $sentence_{j,i}$ stands for a temporary representation of the sentence at time *j,i*. Boxes represent functional organization rather than specific brain regions. Arrows reflect the flow of information: each level of representation is updated incrementally, combining information from the same level at the previous time step (horizontal arrows) and the level below (bottom-up arrows).
(C) The unified architecture implements a unified, global context model through information flowing down the hierarchy, such that expectations at lower levels incorporate information accumulated at the sentence level. Relevant differences from the local context model are in red. Note that while the arrows only cross one level at a time, the information is propagated in steps and eventually crosses all levels.

These hypotheses make different predictions for brain responses sensitive to language statistics. Probabilistic speech representations, as in Figure 1, are linked to brain activity through information theoretic complexity metrics[51]. The most common linking variable is *surprisal*, which is equivalent to the difficulty incurred in updating an incremental representation of the input[4]. A second information theoretic measure that has been found to independently predict brain activity is entropy[45,47], a measure of the uncertainty in a probability distribution. Because entropy is a function of a distribution, entropy differs depending on the unit of classification. This allows distinguishing between the entropy of recognizing the current partial word, and the entropy of predicting the next phoneme (see Methods for details). Entropy might relate to neuronal processes in at least two ways. First, the amount of uncertainty might reflect the amount of competition among different representations, which might play out through a neural process such as lateral inhibition[36]. Second, uncertainty might also be associated with increased sensitivity to bottom-up input, because the input is expected to be more informative[52,53].

Models for responses to continuous speech

To test how context is used in continuous speech processing, we compared the ability of three different context models to predict MEG responses, corresponding to the three levels in Figure 1-B (see Figure 2). The context models all incrementally estimate a probability distribution at each phoneme position, but they differ in the amount and kind of context they incorporate. Throughout, we used n-gram models to estimate sequential dependencies because they are powerful language models that can capture effects of language statistics in a transparent manner, with minimal assumptions about the underlying cognitive architecture[4,5,54].

Sublexical context model: A 5-gram model estimates the prior probability for the next phoneme given the 4 preceding phonemes. This model reflects simple phoneme sequence statistics[42,43] and is unaware of word boundaries. Such a model is thought to play an important role in language acquisition[55–57], but it is unknown whether it has a functional role in adult speech processing. The sublexical model predicted brain responses via the phoneme surprisal and entropy linking variables.

Word context model: This model implements the cohort model of word perception[58], applied to each word in isolation. The first phoneme of the word generates a probability distribution over the lexicon, including all words starting with the given phoneme, and each word's probability proportional to the word's relative unigram frequency. Each subsequent phoneme trims this distribution by removing words that are inconsistent with that phoneme. Like the sublexical model, the lexical model can be used as a predictive model for upcoming phonemes, yielding phoneme surprisal and entropy variables. In addition, the lexical model generates a probability distribution over the lexicon, which yields a cohort entropy variable.

Sentence context model: The sentence model is closely related to the lexical model, but each word's prior probability is estimated from a lexical 5-gram model. While a 5-gram model misses longer-range linguistic dependencies, we use it here as a conservative initial approximation of high level linguistic and interpretive constraints[5]. The sentence model implements cross-hierarchy predictions by using the sentence context in concert with the partial current word to predict upcoming phonemes. Brain activity is predicted from the same three variables as from the word context model.

We evaluated these different context models in terms of their ability to explain held-out MEG responses, and the latency of the brain responses associated with each model. An architecture based on local context models, as in Figure 1-B, predicts a temporal sequence of responses as information passes up the hierarchy, with earlier responses reflecting lower order context models. In contrast, a unified architecture, as in Figure 1-C, predicts that the sentence context model should exhaustively explain brain responses, because all representational levels use priors derived from the sentence context. Finally, architectures that entail multiple kinds of models predict that different context models might explain response components, possibly in different anatomical areas.

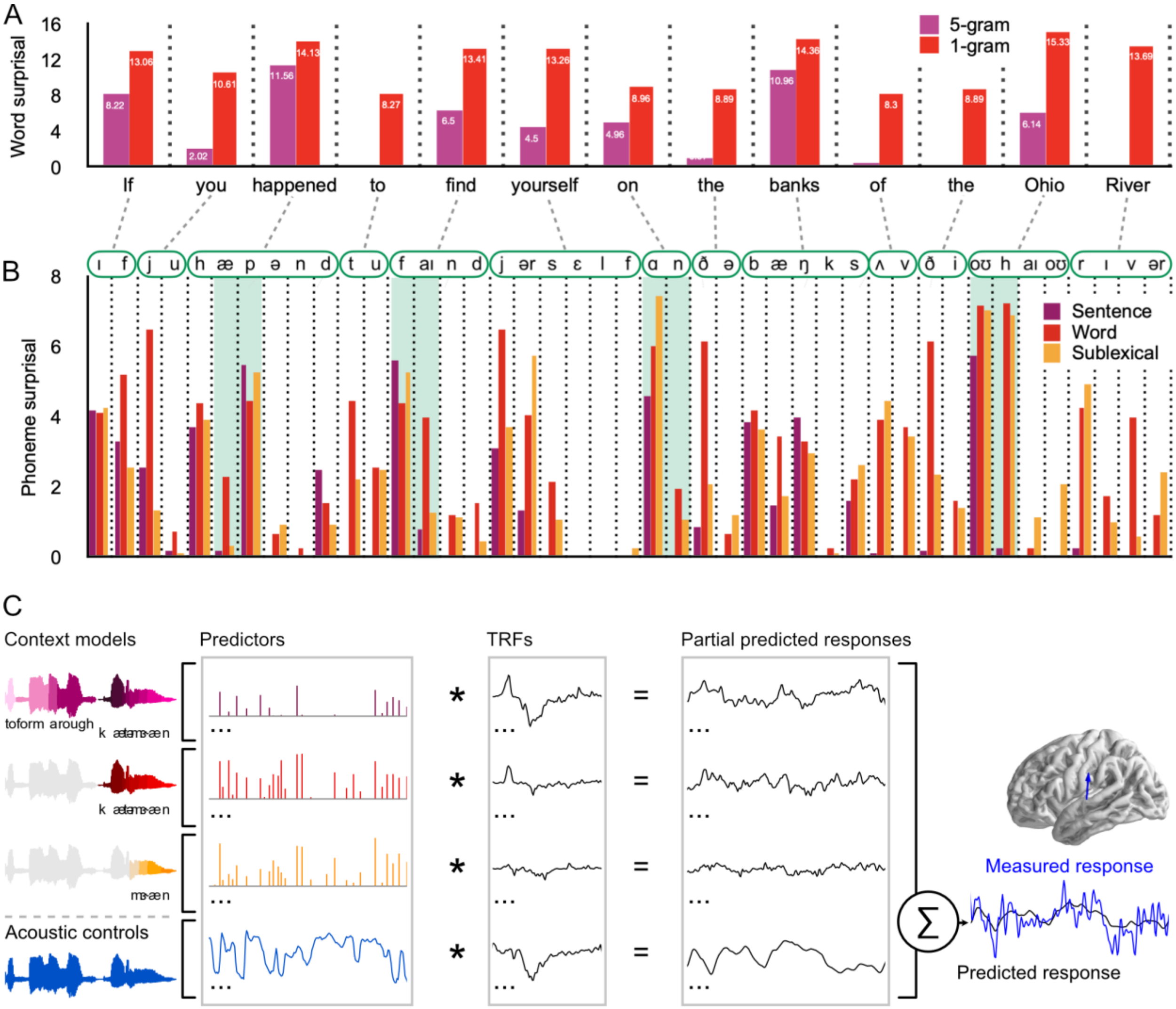

Figure 2. **Models for predictive speech processing based on the sentence, lexical and sublexical context, used to predict MEG data**

(A) Example of word-by-word surprisal. The sentence (5-gram) context generally leads to a reduction of word surprisal, but the magnitude of the reduction differs substantially between words.

(B) Sentence level predictions propagate to phoneme surprisal, but not in a linear fashion. For example, in the word *happened*, the phoneme surprisal based on all three models is relatively low for the second phoneme /æ/ due to the high likelihood of word candidates like *have* and *had*. However, the next phoneme is /p/ and phoneme surprisal is high across all three models. On the other hand, for words like *find*, *on* and *Ohio*, the sentence-constrained phoneme surprisal is disproportionately low for subsequent phonemes, reflecting successful combination of the sentence constraint with the first phoneme.

(C) Phoneme-by-phoneme estimates of information processing demands, based on different context models, were used to predict MEG responses through multivariate temporal response functions (mTRFs)[59]. mTRFs were estimated jointly such that each predictor, convolved with the corresponding TRF, predicted a partial response, and the point-wise sum of partial responses constituted the predicted MEG response. See Methods for details.

Results

Twelve participants listened to ~45 minutes of a nonfiction audiobook. Multivariate temporal response functions (mTRFs) were used to jointly predict held-out, source localized MEG responses (Figure 2-C). To test whether each context model is represented neurally, the predictive power of the full model including all predictors was compared with the predictive power of a model that was estimated without the predictor variables belonging to this specific context model.

Phoneme-, Word- and Sentence-constrained models co-exist in the brain

Each context model significantly improves the prediction of held-out data, even after controlling for acoustic features and the other two context models (Figure 3-A). Each of the three context models' source localization is consistent with sources in the superior temporal gyrus (STG), thought to support phonetic and phonological processing[60]. In addition, the sentence constrained model also extends to more ventral parts of the temporal lobe, consistent with higher-level language processing[61,62]. For comparison, the predictive power of the acoustic features is shown in Figure 3-D. At each level of context, surprisal and entropy contribute about evenly to the model's predictive power (Figure 3-B).

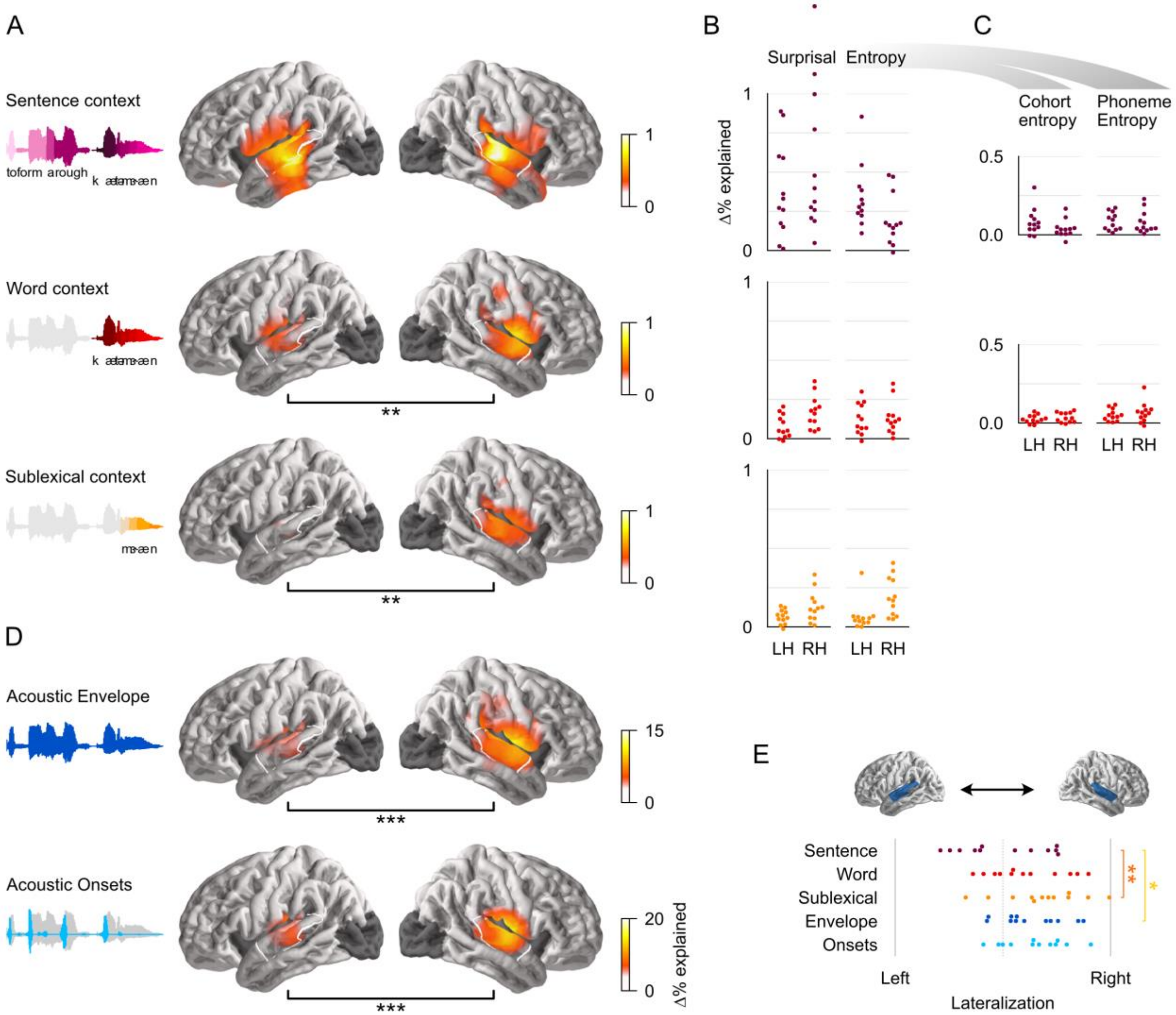

Figure 3. **All context models significantly contribute to predictions of brain responses**

(A) Each context model significantly improves predictions of held-out MEG data in both hemispheres ($t_{\max} \geq 6.16$, $p \leq .005$). Black bars below anatomical plots indicate a significant difference between hemispheres. The white outline indicates a region of interest (ROI) used for measures shown in (B), (C) and (E).

(B) Surprisal and entropy have similar predictive power in each context model (each dot represents one subject; predictive power averaged in the ROI). Cohort- and phoneme entropy are combined here because the predictors are highly correlated and hence share a large portion of their explanatory power. Individual values provided in Supplementary Data. LH: left hemisphere; RH: right hemisphere.

(C) Even when tested individually, excluding variability that is shared between the two, cohort- and phoneme entropy at each level significantly improved predictions. A significant effect of sentence-constrained phoneme entropy is evidence for cross-hierarchy integration.

(D) Predictive power of the acoustic feature representations.

(E) The lateralization index ($LI = R/(L + R)$) indicates that the sublexical context model is more right-lateralized than the sentence context model. Left: *LI* = 1; Right: *LI* = 0. Significance levels: * *p* ≤ .05; ** *p* ≤ .01; *** *p* ≤ .001.

The predictive power of the local context models is inconsistent with the hypothesis of a single, unified context model (Figure 1-C). Instead, it suggests that different neural representations incorporate different kinds of context. We next pursued the question of how these different representations are organized hierarchically. While surprisal depends on the conditional probability of a discrete event and is agnostic to the underlying unit of representation[4,5], entropy depends on the units over which probabilities are calculated. Entropy can thus potentially distinguish between whether brain responses reflect uncertainty over the next phoneme, or uncertainty over the word currently being perceived. This distinction is particularly interesting for the sentence context model: if predictions are constrained to using context within a hierarchical level, as in Figure 1-B, then the sentence context should affect uncertainty about the upcoming word, but not uncertainty about the upcoming phoneme. On the other hand, a brain response related to sentence-conditional phoneme entropy would constitute evidence for cross-hierarchy predictions, with sentence level information predicting upcoming phonemes.

Even though phoneme and cohort entropy were highly correlated (sentence context: *r* = .92; word context: *r* = .90), each of the four representations was able to explain variability in the MEG responses that could not be attributed to any of the other representations (Figure 3-C; all t_{11} ≥ 2.49, *p* ≤ .030). This suggests that the sentence context model is not restricted to predicting upcoming words, but also generates expectations for upcoming phonemes. This is thus evidence for cross-hierarchy top-down information flow, indicative of a unified language model that aligns representations across hierarchical levels. Together, these results thus indicate that the brain does maintain a unified context model, but that it *also* maintains more local context models.

Different context models affect different neural processes

All three context models individually contribute to neural representations, but are these representations functionally separable? While all three context models improve predictions in both hemispheres, the sentence constrained model does so symmetrically, whereas the lexical and sublexical models are both more powerful in the right hemisphere than in the left hemisphere (Figure 3-A). The sublexical context model is indeed significantly more right-lateralized than the sentence model (t_{11} = 4.33, *p* = .001; Figure 3-E), while the word model is only numerically more right-lateralized than the sentence model (t_{11} = 1.48, *p* = .167). This difference in lateralization suggests some anatomical differentiation in the representations of different context models, with the left hemisphere primarily relying on a unified model of the sentence context, and the right hemisphere more broadly keeping track of different context levels.

Given that all three context models are represented in the STG, especially in the right hemisphere, a separate question concerns whether, within a hemisphere, the different context models predict activity in the same or different neural sources. While MEG source localization does not allow precisely separating different sources in close proximity, it does allow statistically testing whether two effects originate from the same or from a different configuration of neural sources[63]. The null hypothesis of such a test[64] is that a single neural process, corresponding to a fixed configuration of current sources, generates activity that is correlated with all three context models. The alternative hypothesis suggests some differentiation between the configuration of sources recruited by the different models. Results indicate that, in the right hemisphere, all three context models, as well as the two acoustic models, originate from different source configurations ($F_{(175, 1925)} \geq 1.25$, $p \leq .017$). In the left hemisphere, the sentence constrained model is localized differently from all other models ($F_{(179, 1969)} \geq 1.38$, $p < .001$), whereas there is no significant distinction among the other models (possibly due to lower power due to the weaker effects in the left hemisphere for all but the sentence model). In sum, these results suggest that the different context models are maintained by at least partially separable neural processes.

Sentence context affects early responses and dominates late responses

The TRFs estimated for the full model quantify the influence of each predictor variable on brain responses over a range of latencies (Figure 2-C). Figure 4 shows the response magnitude to each predictor variable as a function of time, relative to phoneme onset. For an even comparison between predictors, TRFs were summed in the anatomical region in which any context model significantly improved predictions. Note that responses prior to 0 ms are plausible due to coarticulation, by which information about a phoneme's identity can already be present in the acoustic signal prior to the conventional phoneme onset[65,66]. Figure 5 shows the anatomical distribution of responses related to the different levels of context.

Surprisal quantifies the incremental update to a context model due to new input. A brain response related to surprisal therefore indicates that the input is brought to bear on a neural representation that uses the corresponding context model. Consequently, the latencies of brain responses related to different context models are indicative of the underlying processing architecture. In an architecture in which information is sequentially passed to higher level representations with broadening context models (Figure 1-B), responses should form a temporal sequence from narrower to broader contexts. However, in contrast to this prediction, the observed responses to surprisal suggest that bottom-up information reaches representations using sentence- and word-level contexts *simultaneously* at an early response peak (Figure 4-A; sentence: 78 ms, SD = 24 ms; word: 76 ms, SD = 11 ms). Sublexical surprisal is associated with a lower response magnitude overall, but also exhibits an early peak at 94 ms (SD = 26 ms). This suggests a parallel processing architecture in which different context representations are activated simultaneously by new input. Later in the timecourse the responses dissociate more strongly, with a large, extended response reflecting the sentence context, but not the word context starting at around 205 ms ($t_{max} = 5.27$, $p = .007$). The lateralization of the TRFs is consistent with the trend observed for predictive power: a symmetric response reflecting the unified sentence context, and more right-lateralized responses reflecting the more local contexts (Figure 4-B).

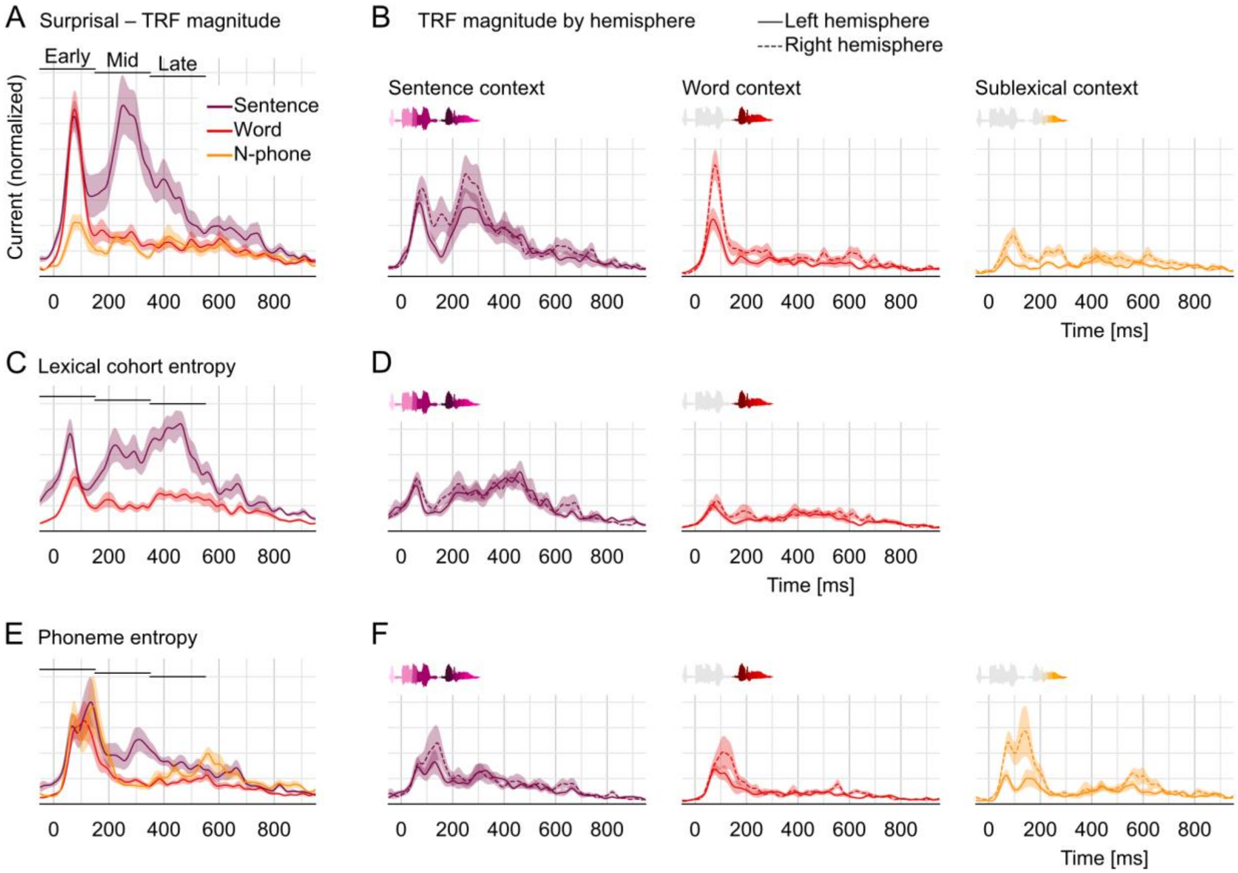

Figure 4. **Early responses reflect parallel activation of all context models, later responses selectively reflect activity in the sentence-constrained model**

(A) Current magnitude of TRFs to phoneme surprisal for each level of context (mean and within-subject standard error[67]; y-axis scale identical in all panels of the figure). Bars indicate time windows corresponding to source localizations shown in Figure 5.

(B) When plotted separately for each hemisphere, relative lateralization of the TRFs is consistent with the lateralization of predictive power (Figure 3).

(C-D) TRFs to lexical cohort entropy are dominated by the sentence context model.

(E-F) TRFs to phoneme entropy are similar between context models, consistent with parallel use of different contexts in predictive models for upcoming speech.

Sentence context dominates word recognition, all contexts drive phoneme predictions

Brain responses related to entropy indicate that neural processes are sensitive to uncertainty or competition in the interpretation of the speech input. Like surprisal, such a response suggests that the information has reached a representation that has incorporated the corresponding context. In addition, because entropy measures uncertainty regarding a categorization decision, the response to entropy can distinguish between different levels of categorization: uncertainty about the current word (cohort entropy) versus uncertainty about the next phoneme (phoneme entropy).

The TRFs to cohort entropy suggest a similar pattern as those to surprisal (Figure 4 C-D). Both cohort representations are associated with an early peak (sentence context: 56 ms, SD = 28 ms;

word context: 80 ms, SD = 23 ms), followed only in the sentence constrained cohort by a later sustained effect. In contrast to surprisal, however, even early responses to cohort-entropy are dominated by the sentence context (t_{max} = 5.35, p = .004 at 43 ms; later responses: t_{max} = 7.85, p < .001 at 461 ms). This suggests that lexical representations are overall most strongly activated in a model that incorporates the sentence context.

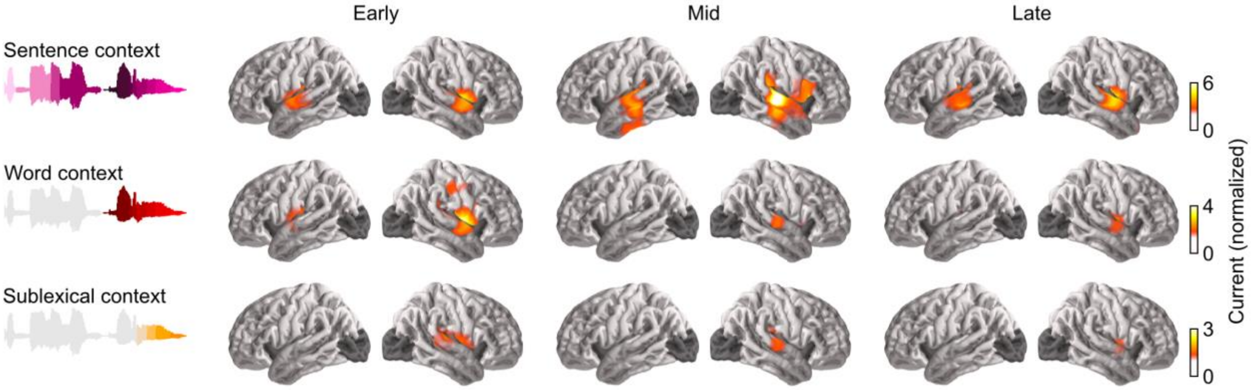

Figure 5. **All context models engage the superior temporal gyrus at early response, mid-latency responses incorporating the sentence context also engage more ventral temporal areas**
Current magnitude associated with different levels of context representing early (-50 – 150 ms), mid-latency (150 – 350 ms) and late (350 – 550 ms) responses. The color-scale is adjusted for different predictors to avoid images dominated by the spatial dispersion characteristic of MEG source estimates.

In contrast to surprisal and cohort entropy, the responses to phoneme entropy are similar for all levels of context, dominated by an early and somewhat broader peak (Figure 4 E-F). There is still some indication of a second, later peak in the response to sentence-constrained phoneme entropy, but this might be due to the high correlation between cohort and phoneme entropy. A direct comparison of sentence-constrained cohort and phoneme entropy indicates that early processing is biased towards phoneme entropy (though not significantly) while later processing is biased towards cohort entropy (t_{max} = 4.74, p = .017 at 231 ms).

In sum, the entropy results suggest that all context representations drive a predictive model for upcoming phonemes. This is reflected in a short-lived response in STG, consistent with the fast rate of phonetic information. Simultaneously, the incoming information is used to constrain the cohort of word candidates matching the current input, with lexical activations primarily driven by a unified model that incorporates the sentence context.

Mid-latency, sentence-constrained processing engages larger parts of the temporal lobe

Source localization suggests that early activity originates from the vicinity of the auditory cortex in the upper STG, regardless of context (Figure 5). The precise source configuration in the right STG nevertheless differs between contexts in the early time window (sentence vs word: $F_{(175,\ 1925)}$ = 2.08, p < .001; word vs sublexical: $F_{(175,\ 1925)}$ = 5.99, p < .001). More notably, the sentence-based responses in the mid-latency window recruits more sources, localized to the middle and inferior temporal lobe. Accordingly, the sentence-based responses in the mid-latency window differs significantly from the early window (left hemisphere (L): $F_{(179,\ 1969)}$ = 1.72, p < .001; right hemisphere (R): $F_{(175,\ 1925)}$ = 5.48, p < .001). These results suggest that phonetic information initially engages a set of sources in the STG, while a secondary stage then engages more ventral sources that specifically represent the sentence context.

No evidence for a trade-off between contexts

We interpret our results as evidence that different context models are maintained in parallel. An alternative possibility is that there is some trade-off between contexts used, and it only appears in the averaged data as if all models were operating simultaneously. This alternative predicts a negative correlation between the context models, reflecting the trade-off in their activation. No evidence was found for such a trade-off, as correlation between context models were generally neutral or positive across subjects and across time (see Supplementary Figure 1).

Discussion

The present MEG data provide clear evidence for the existence of a neural representation of speech that is unified across representational hierarchies. This representation incrementally integrates phonetic input with information from the multi-word context within about 100 ms. However, in addition to this globally unified representation, brain responses also show evidence of separate neural representations that use more local contexts to process the same input.

Parallel representations of speech using different levels of context

The evidence for a unified global model suggests that there is a functional brain system that processes incoming phonemes while building a representation that incorporates constraints from the multi-word context. A possible architecture for such a system is the one shown in Figure 1-C, in which a probabilistic representation of the lexical cohort mediates between sentence and phoneme level representations: the sentence context modifies the prior expectation for each word, which is in turn used to make low-level predictions about the phonetic input. While there are different possible implementations for such a system, the key feature is that the global sentence context is used to make predictions for and interpret low-level phonetic, possibly even acoustic[68] input.

A second key result from this study, however, is evidence that this unified model is not the only representation of speech. Brain responses also exhibited evidence for two other, separate functional systems that process incoming phonemes while building representations that incorporate different, more constrained kinds of context: one based on a local word context, processing the current word with a prior based on context-independent lexical frequencies, and another based on the local phoneme sequence regardless of word boundaries. Each of these three functional systems generates its own predictions for upcoming phonemes, resulting in parallel responses to phoneme entropy. Each system is updated incrementally at the phoneme rate, reflected in early responses to surprisal. However, each system engages an at least partially different configuration of neural sources, as evidenced by the localization results.

Together, these results suggest that multiple predictive models process speech input in parallel. An architecture consistent with these observations is sketched in Figure 6: three different neural systems receive the speech input in parallel. Each representation is updated incrementally by arriving phonemes. However, the three systems differ in the extent and kind of context that they incorporate, each generating its own probabilistic beliefs about the current word and/or future phonemes. For instance, the sublexical model uses the local phoneme history to predict upcoming phonemes. The incrementality of the updates is reflected in the inputs to the sublexical model at time *k*+1, combining the state of the sublexical model at time k and the phoneme input from time *k*. The same incremental update pattern applies to the word and sentence models.

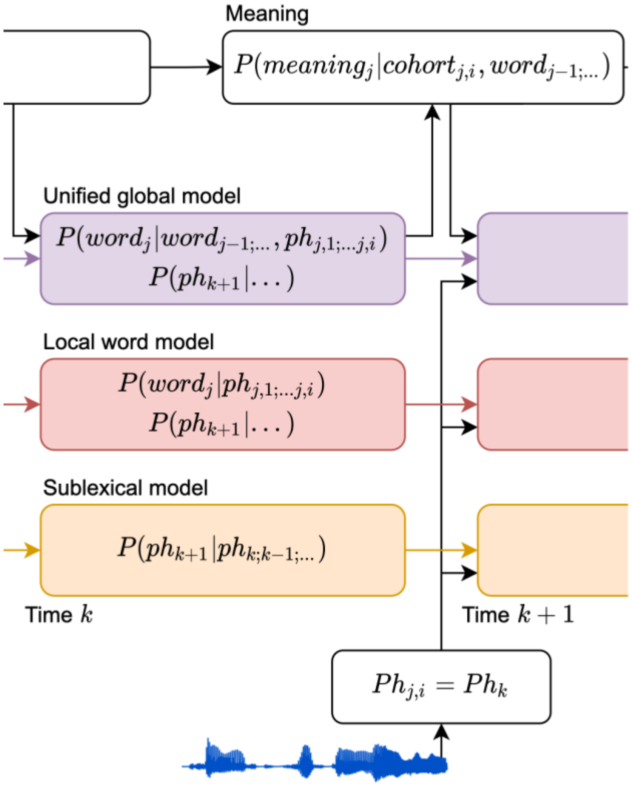
Meaning
$P(meaning_j|cohort_{j,i}, word_{j-1;...})$
Unified global model
$P(word_j|word_{j-1;...}, ph_{j,1;...j,i})$
$P(ph_{k+1}|...)$
Local word model
$P(word_j|ph_{j,1;...j,i})$
$P(ph_{k+1}|...)$
Sublexical model
$P(ph_{k+1}|ph_{k;k-1;...})$
Time k
Time $k+1$
$Ph_{j,i} = Ph_k$

Figure 6. **An architecture for speech perception with multiple parallel context models**

A model of information flow, consistent with brain signals reported here. Brain responses associated with Information theoretic variables provided separate evidence for each of the probability distributions in the colored boxes. From left to right, the three different context models (sentence, lexical and sublexical) update incrementally as each phoneme arrives. The cost of these updates is reflected in the brain response related to surprisal. Representations also include probabilistic representations of words and upcoming phonemes, reflected in brain responses related to entropy.

A listener whose goal is comprehending a discourse-level message might be expected to rely primarily on the unified, sentence constrained context model. Consistent with this, there is some evidence that this model has a privileged status. Among the linguistic models, the unified model has the most explanatory power and clearly bilateral representation (Figure 3). In addition, while activity in local models was short-lived, the unified model was associated with extended activation for up to 600 ms and recruitment of more ventral regions of the temporal lobe (Figure 4 and 5). This suggests that the update in the unified model is normally more extensive than the local models, and could indicate that the unified model most commonly drives semantic as well as form representations, while the short-lived local models might be restricted to form-based representations.

Implications for speech processing

A longstanding puzzle in the comprehension literature has been why activation of non-contextually supported candidates is sometimes reported[38,39], if top-down sentence context rapidly feeds down to early levels of speech perception. Parallel activation of lexical candidates based on sentence and word context models can explain these findings. Short-lived brain responses (up to 150 ms after phoneme onset) show evidence of parallel activation of sentence-constrained as well as sentence-independent word candidates. The co-existence of these two candidate sets can explain short-lived priming of sentence-inappropriate candidates. Whereas brain responses related to sentence-independent candidates are transient, brain responses related to sentence-appropriate candidates exhibit a secondary, sustained response (150-550 ms), explaining selective priming of sentence-appropriate candidates at longer delays.

If context-constrained candidates are immediately available, then why maintain alternative, sentence-independent candidates or even sublexical probabilistic phoneme sequences? One functional advantage might be faster recovery when sentence-based predictions turn out to be misleading. Such an effect has been described in reading, where contextually unexpected continuations are not associated with a proportional increase in processing cost[69,70].

Similarly, a representation of sublexical phoneme sequences might be functionally relevant when encountering input that is noisy or not yet part of the lexicon. Phoneme transition probabilities are generally higher within words than between words, such that low probability phoneme transitions are cues to word boundaries[56,71]. Statistical phoneme sequence models might thus play an important role in language acquisition by bootstrapping lexical segmentation of continuous speech[55–57]. Even in adult speech perception, they might have a similar function when encountering novel words, such as domain-specific vocabularies or personal names[27]. Finally, the linguistic context can be highly informative for phoneme recognition[72], and different levels of context might make complementary contributions.

The parallel model suggested in Figure 6 has a special theoretical appeal over the two-stage explanation: Bayesian accounts of perception suggest that listeners generate a prior, reflecting an estimate of future input, and compare this prior to the actual input to compute a posterior probability, or interpretation of the sensory percept. In architectures that allow different priors at sequential hierarchical levels (such as Figure 1-B), higher levels receive the posterior interpretation of the input from the lower levels, rather than the unbiased input itself. This is suboptimal when considering a Bayesian model of perception, because the prior of lower levels is allowed to distort the bottom-up evidence before it is compared to the prior generated by higher levels[73]. In contrast, the parallel representations favored by the evidence presented here allows undistorted bottom-up information to be directly compared with the context model for each definition of context. The parallel model can thus explain empirical effects of local context priors while avoiding this theoretical problem associated with sequential models.

Evidence for graded linguistic predictions

There is broad agreement that language processing involves prediction, but the exact nature of these predictions is more controversial[74–78]. Much of the debate is about whether humans can represent distributions over many likely items, or just predict specific items. Previous research showing an early influence of sentence context on speech processing[7–9,79] has typically relied on specifically designed, highly constraining contexts which are highly predictive of a specific lexical item. In such highly predictive contexts, listeners might indeed predict specific items, and such predictions might be linked to the left-lateralized speech productions system[44,77]. However, such a mechanism would be less useful in more representative language samples, in which highly predictable words are rare[69]. In such situations of limited predictability, reading time data suggest that readers instead make graded predictions, over a large number of possible continuations[5,69]. Alternatively, it has been suggested that what looks like graded predictions could actually be pre-activation of specific higher-level semantic and syntactic features shared among the likely items[69,77,80–82], without involving prediction of form-based representations. The present results, showing brain responses reflecting sentence-constrained cohort- and phoneme entropy, provide a new kind of evidence in favor of graded probabilistic predictions, involving predictive representations at least down to the phoneme level.

Bilateral pathways to speech comprehension

Our results suggest that lexical/phonetic processing is largely bilateral. This is consistent with extensive clinical evidence for bilateral receptive language ability[83,84,61], and suggestions that the right hemisphere might even play a distinct role in complex, real-world language processing[85,86]. In healthy participants, functional lateralization of sentence processing has been studied using visual half-field presentation[87]. Overwhelmingly, results from these studies suggest that lexical processing in both hemispheres is dominated by sentence meaning[87–90]. This is consistent with the strong bilateral representation of the unified model of speech found here. As in the visual studies, the similarity of the response latencies in the two hemispheres implies that right-hemispheric effects are unlikely to be due to inter-hemispheric transfer from the left hemisphere (Figure 4).

Nevertheless, response patterns are not identical between hemispheres. Hemispheric differences in visual half-field studies have been interpreted as indicating that the left hemisphere processes language in a maximally context-sensitive manner, whereas the right hemisphere is more biased towards a bottom-up interpretation of sensory input[44]. Our results suggest a modification of this proposal, indicating that both hemispheres rely on sentence-based graded predictions, but that the right hemisphere *additionally* maintains stronger representations of local contexts. Finally, lateralization might also depend on task characteristics such as stimulus familiarity[45], and in highly constraining contexts the left hemisphere might engage the language production system to make specific predictions[44,77].

Conclusions

Prior research on the use of context during language processing has often focused on binary distinctions, such as asking whether context is or is not used to predict future input. Such questions assumed a single serial or cascaded processing stream. Here we show that this assumption might have been misleading, because different predictive models are maintained in parallel. Our results suggest that robust speech processing is based on probabilistic predictions using different context models in parallel and cutting across hierarchical levels of representations.

Acknowledgements

Work supported by a UMD BBI seed grant to EFL and JZS; NSF BCS-1749407 to EFL; NIH R01DC014085 & NSF SMA-1734892 to JZS; NSF BCS-1754284; ONR MURI Award N00014–18–1–2670 to PSR.

Materials and Methods

Participants

Twelve native speakers of English were recruited from the University of Maryland community (6 female, 6 male, age mean = 21 years, range 19-23). None reported any neurological or hearing impairment. According to self-report using the Edinburgh Handedness Inventory[91], 11 were right-handed and one left-handed. All subjects provided informed consent in accordance with the University of Maryland Institutional Review Board. Subjects either received course credit (n=4) or were paid for their participation (n=8).

Stimuli

Stimuli consisted in eleven excerpts from the audiobook version of *The Botany of Desire* by *Michael Pollan*[92]. Each excerpt was between 210 and 332 seconds long, for a total of 46 minutes and 44 seconds. Excerpts were selected to create a coherent narrative and were presented in chronological order to maximize deep processing for meaning.

Procedure

During MEG data acquisition, participants lay in a supine position. They were allowed to keep their eyes open or closed to maximize comfort. Stimuli were delivered through foam pad earphones inserted into the ear canal at a comfortably loud listening level. After each segment, participants answered 2-3 questions relating to its content and had an opportunity to take a short break.

Data acquisition and preprocessing

Brain responses were recorded with a 157 axial gradiometer whole head MEG system (KIT, Kanazawa, Japan) inside a magnetically shielded room (Vacuumschmelze GmbH & Co. KG, Hanau, Germany) at the University of Maryland, College Park. Sensors (15.5 mm diameter) are uniformly distributed inside a liquid-He dewar, spaced ~25 mm apart, and configured as first-order axial gradiometers with 50 mm separation and sensitivity better than 5 $\mathrm{fT \cdot Hz^{-1/2}}$ in the white noise region (> 1 KHz). Data were recorded with an online 200 Hz low-pass filter and a 60 Hz notch filter at a sampling rate of 1 kHz.

Recordings were pre-processed using mne-python[93]. Flat channels were automatically detected and excluded. Extraneous artifacts were removed with temporal signal space separation[94]. Data were filtered between 1 and 40 Hz with a zero-phase FIR filter (mne-python 0.20 default settings). Extended infomax independent component analysis[95] was then used to remove ocular and cardiac artifacts. Responses time-locked to the speech stimuli were extracted, low pass filtered at 20 Hz and resampled to 100 Hz.

Five marker coils attached to participants' head served to localize the head position with respect to the MEG sensors. Head position was measured at the beginning and at the end of the recording session and the two measurements were averaged. The FreeSurfer[96] ''fsaverage'' template brain was coregistered to each participant's digitized head shape (Polhemus 3SPACE FASTRAK) using rotation, translation, and uniform scaling. A source space was generated using four-fold icosahedral subdivision of the white matter surface, with source dipoles oriented perpendicularly to the cortical surface. Regularized minimum $\ell 2$ norm current estimates[97,98] were computed for all data using an empty room noise covariance (λ = 1/6). The temporal response function analysis was restricted to brain areas of interest by excluding the occipital lobe, insula and midline structures based on the "aparc" FreeSurfer parcellation[99]. Excluded areas are shaded gray in Figure 3. A preliminary analysis (see below) was restricted to the temporal lobe (superior, middle and inferior temporal gyri, Heschl's gyrus and superior temporal sulcus).

Predictor variables

Acoustic model

To control for brain responses to acoustic features, all models included an 8 band gammatone spectrogram and an 8 band acoustic onset spectrogram[100], both covering frequencies from 20 to 5000 Hz in equivalent rectangular bandwidth (ERB) space[101] and scaled with exponent 0.6[102].

Word- and phoneme segmentation

A pronunciation dictionary was generated by combining the Carnegie-Mellon University pronunciation dictionary with the Montreal Forced Aligner[103] dictionary and adding any additional words that occurred in the stimuli. Transcripts were then aligned to the acoustic stimuli using the Montreal Forced Aligner [103] version 1.0.1. All models included control predictors for word onsets (equal value impulse at the onset of each word) and phoneme onsets (equal value impulse at the onset of each non-word initial phoneme).

Context-based predictors

All experimental predictor variables consistent of one value for each phoneme and were represented as a sequence of impulses at all phoneme onsets. The specific values were derived from three different linguistic context models.

Sublexical context model

The complete SUBTLEX-US corpus[104] was transcribed by substituting the pronunciation for each word and concatenating those pronunciations across word boundaries (i.e., no silence between words). Each line was kept separate since lines are unordered in the SUBTLEX corpus. The resulting phoneme sequences were then used to train a 5-gram model using KenLM[105]. This 5-gram model was then used to derive phoneme surprisal and entropy.

The surprisal of experiencing phoneme ph_k at time point k is inversely related to the likelihood of that phoneme, conditional on the context (measured in bits): $I(ph_k) = -log_2(p(ph_k|context))$. In the case of the 5-phone model this context consists of the preceding 4 phonemes, $ph_{k-4;\dots k-1}$.

The entropy H (Greek Eta) at phoneme position ph_k reflects the uncertainty of what the next phoneme, ph_{k+1} will be. It is defined as the expected (average) surprisal at the next phoneme, $H(ph_k) = -\sum_{ph}^{phonemes} p(ph_{k+1} = ph|context) \log_2(p(ph_{k+1} = ph|context))$. Based on the 5-phone model, the context here is $ph_{k-3;\dots k}$.

Lexical context model

The lexical context model takes into account information from all phonemes that are in the same word as, and precede the current phoneme[45] and is based on the cohort model of word perception[58]. At word onset, the prior for each word is proportional to its frequency in the Corpus of Contemporary American English (COCA)[106]. With each subsequent phoneme, the probability for words that are inconsistent with that phoneme is set to 0, and the remaining distribution is renormalized. Phoneme surprisal and entropy are then calculated as above, but with the context being all phonemes in the current word so far. I addition, lexical entropy is calculated at each phoneme position as the entropy in the distribution of the cohort $H(ph_{j,i}) = -\sum_{word}^{lexicon} p(word_j = word|context)\log_2(p(word_j = word|context))$ where *j* is the index of the word, *i* is the index of the current phoneme within word *j*, and the context consists of phonemes $ph_{j,1;\ldots j,i-1}$.

Sentence context model

The sentence context model was implemented like the lexical context model, but with the addition of lexical priors based on the 5-gram word context. A 5-gram model was trained on COCA[106] with KenLM[105]. Then, at the onset of each word, the cohort was initialized with each word's prior set to its probability given the 4 preceding words in the 5-gram model.

Reverse correlation

Multivariate temporal response functions (mTRFs) were computed independently for each subject and each virtual current source[59,107]. The neural response at time t, y_t was predicted jointly from N predictor time series $x_{i,t}$ convolved with a corresponding mTRF $h_{i,\tau}$ of length T:

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

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

For estimation using 4-fold cross-validation, each subject's data were concatenated along the time axis and split into 4 contiguous segments of equal length. The mTRFs for predicting the responses in each segment were trained on the remaining 3 segments. Each of the 4 training runs in turn consisted of 3 iterations, in which the 3 segments were divided into 2 training segments and 1 validation segment. In each training run, an mTRF was estimated using an iterative coordinate descent algorithm[108] to minimize the $\ell 1$ error. The mTRF was iteratively modified based on the maximum error reduction in the training set (the steepest coordinate descent) and validated based on the error in the validation set. Whenever a training step caused an increase of error in the validation set, the TRF for the predictor responsible for the increase was frozen, and training continued until the whole mTRF was frozen. The 3 mTRFs from the 3 training runs were then averaged to predict responses in the left-out testing segment.

Model comparisons

Model quality was quantified through the $\ell 1$ norm of the residuals. For this purpose, the predicted responses for the 4 test segments, each based on mTRFs estimated on the other 3 segments, were concatenated again. To compare the predictive power of two models, the difference in the residuals of the two models was calculated at each virtual source dipole. This difference map was smoothed (Gaussian window, SD = 5 mm) and tested for significance using a mass-univariate one-sample t-test with threshold-free cluster enhancement (TFCE)[109] and a null distribution based on the full set of 4095 possible permutations of the 12 difference maps. For effect size comparison we report t_{max}, the largest t-value in the significant ($p \leq .05$) area.

The full model consisted of the following predictors: acoustic spectrogram (8 bands); acoustic onset spectrogram (8 bands); word onsets; phoneme onsets; sublexical context model (phoneme surprisal and phoneme entropy); lexical context model (phoneme surprisal, phoneme entropy and word entropy); sentence context model (phoneme surprisal, phoneme entropy and word entropy).

For each of the tests reported in Figure 3, mTRFs were re-estimated using a corresponding subset of the predictors in the full model. For instance, to calculate the predictive power for a given level of context, the model was re-fit using all predictors except the predictors of the level under investigation. Each plot thus reflects the variability that can *only* be explained by the level in question. This is generally a conservative estimate for the predictive power because it discounts any explanatory power based on variability that is shared with other predictors.

To express model fits in a meaningful unit, the explainable variability was estimated through the largest possible explanatory power of the full model (maximum across the brain of the measured response minus residuals, averaged across subjects). All model fits were then expressed as % of this value. For visualization, brain maps are not masked by significance to accurately portray the continuous nature of MEG source estimates.

ROI

To allow for univariate analyses of predictive power, an ROI was used including a region responsive to all context models (white outline in Figure 3-A). This ROI was defined as the posterior 2/3 of the combined Heschl's gyrus and STG "aparc" label, separately in each hemisphere.

Tests of lateralization

For spatio-temporal tests of lateralization (Figure 3-A and D) the difference map was first morphed to the symmetric "fsaverage_sym" brain[110], and the data from the right hemisphere was morphed to the left hemisphere. Once in this common space, a mass-univariate repeated measures t-test with TFCE was used to compare the difference map from the left and right hemisphere.

Tests of localization difference

A direct comparison of two localization maps can have misleading results due to cancellation between different current sources[63] as well as the continuous nature of MEG source estimates[111]. However, a test of localization difference is possible due to the additive nature of current sources[64]. Specifically, for a linear inverse solver as used here, if the relative amplitude of a configuration of current sources is held constant, the topography of the resulting source localization is also unchanged. Consequently, we employed a test of localization difference that has the null hypothesis that the topography of two effect in source space is the same[64]. Localization tests were generally restricted to an area encompassing the major activation seen in Figure 3, based on "aparc" labels[99]: the posterior 2/3 of the superior temporal gyrus and Heschl's gyrus combined, the superior temporal sulcus, and the middle 3/5 of the middle temporal gyrus. For each map, the values in this area were extracted and z-scored (separately for each hemisphere). For each comparison, the two z-scored maps were subtracted, and the resulting difference map was analyzed with a one-way repeated measures ANOVA with factor source location (left hemisphere: 180 sources; right hemisphere: 176 sources). According to the null hypothesis, the two maps should be (statistically) equal, and the difference map should only contain noise. In contrast, a significant effect of source location would indicate that the difference map reflects a difference in topography that is systematic between subjects.

TRF analysis

For the analysis of the TRFs, all 12 mTRFs estimated for each subject were averaged (4 test segments * 3 training runs). TRFs were analyzed in the normalized scale that was used for model estimation.

TRF time-course

To extract the time course of response functions, an ROI was generated including all virtual current sources for which at least one of the three context models significantly improved the response predictions. To allow a fair comparison between hemispheres, the ROI was made symmetric by morphing it to the “fsaverage_sym” brain[110] and taking the union of the two hemispheres. With this ROI, the magnitude of the TRFs at each time point was then extracted as the sum of the absolute current values across source dipoles. These time courses were resampled to 1000 Hz. Peak times were determined by finding the maximum value within a given window for each subject. Time-courses were statistically compared using mass-univariate related measures *t*-tests, with a null distribution based on the maximum statistic in the 4095 permutations (no cluster enhancement).

TRF-localization

To analyze TRF localization, TRF magnitude was quantified as the summed absolute current values in three time-windows, representing early (-50 – 150 ms), mid-latency (150 – 350 ms) and late (350 – 550 ms) responses (see Figure 5). Maps were smoothed (Gaussian window, SD = 5 mm) and tested for localization differences with the same procedure as described above (Tests of localization difference).

Analysis of trade-off between context models

Several analyses were performed to detect a trade-off between the use of the different context models.

Trade-off by subject

One possible trade-off is between subjects: some subjects might rely on sentence context more than local models, whereas other subjects might rely more on local models. For example, for lexical processing, this hypothesis would predict that for a subject for whom the sentence context model is more predictive, the lexical context model should be less and vice versa. According to this hypothesis, the predictive power of the different context models should be negatively correlated across subjects. To evaluate this, we correlations between the predictive power of the different models in the in the mid/posterior STG ROI (see Supplementary Figure 1-A).

Trade-off over time

A second possible trade-off is across time: subjects might change their response characteristics over time to change the extent to which they rely on lower- or higher-level context. For example, the depth of processing of meaningful speech might fluctuate with the mental state of alertness. According to this hypothesis, the predictive power of the different context models should be anti-correlated over time. To evaluate this, we calculated the residuals for the different model fits for each time point, $res_t = abs(y_t - \hat{y}_t)$, aggregating by taking the mean in the mid/posterior STG ROI (separately or each subject). The predictive power was calculated for each model by subtracting the residuals of the model from the absolute values of the measured data (i.e., the residuals of a null model without any predictor). The predictive power for each level of context was then computed by subtracting the predictive power of a corresponding reduced model, lacking the given level of context, from the predictive power of the full model. Finally, to reduce the number of data points the predictive power was summed in 1 second bins.

For each subject, the trade-off between each pair of contexts was quantified as the partial correlation[112] between the predictive power of the two contexts, controlling for the predictive power of the full model (to control for MEG signal quality fluctuations over time). To test for a significant trad-off, a one-sample *t*-test was used for each pair and in each hemisphere, with the null hypothesis that the correlation between contexts over time is 0 (see Supplementary Figure 1-B).

References

1. Christiansen, M.H., and Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. Behav. Brain Sci. *39*, e62.
2. Ferreira, F., and Chantavarin, S. (2018). Integration and Prediction in Language Processing: A Synthesis of Old and New. Curr. Dir. Psychol. Sci. *27*, 443–448.
3. Hale, J. (2003). The Information Conveyed by Words in Sentences. J. Psycholinguist. Res. *32*, 101–123.
4. Levy, R. (2008). Expectation-based syntactic comprehension. Cognition *106*, 1126–1177.
5. Smith, N.J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. Cognition *128*, 302–319.
6. Marslen-Wilson, W., and Tyler, L.K. (1975). Processing structure of sentence perception. Nature *257*, 784–786.
7. Holcomb, P.J., and Neville, H.J. (1991). Natural speech processing: An analysis using event-related brain potentials. Psychobiology *19*, 286–300.
8. Connolly, J.F., and Phillips, N.A. (1994). Event-Related Potential Components Reflect Phonological and Semantic Processing of the Terminal Word of Spoken Sentences. J. Cogn. Neurosci. *6*, 256–266.
9. Van Petten, C., Coulson, S., Rubin, S., Plante, E., and Parks, M. (1999). Time course of word identification and semantic integration in spoken language. J. Exp. Psychol. Learn. Mem. Cogn. *25*, 394–417.
10. Diaz, M.T., and Swaab, T.Y. (2007). Electrophysiological differentiation of phonological and semantic integration in word and sentence contexts. Brain Res. *1146*, 85–100.
11. Broderick, M.P., Anderson, A.J., Liberto, G.M.D., Crosse, M.J., and Lalor, E.C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. Curr. Biol. *28*, 803-809.e3.
12. Willems, R.M., Frank, S.L., Nijhof, A.D., Hagoort, P., and van den Bosch, A. (2016). Prediction During Natural Language Comprehension. Cereb. Cortex *26*, 2506–2516.
13. Weissbart, H., Kandylaki, K.D., and Reichenbach, T. (2020). Cortical Tracking of Surprisal during Continuous Speech Comprehension. J. Cogn. Neurosci. *32*, 155–166.

14. Schmitt, L.-M., Erb, J., Tune, S., Rysop, A., Hartwigsen, G., and Obleser, J. (2020). Predicting speech from a cortical hierarchy of event-based timescales. bioRxiv, 2020.12.19.423616.
15. van Berkum, J.J.A., Zwitserlood, P., Hagoort, P., and Brown, C.M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. Cogn. Brain Res. *17*, 701–718.
16. Nieuwland, M.S., and Van Berkum, J.J.A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. J Cogn Neurosci *18*, 1098–111.
17. Gagnepain, P., Henson, R.N., and Davis, M.H. (2012). Temporal Predictive Codes for Spoken Words in Auditory Cortex. Curr. Biol. *22*, 615–621.
18. Ettinger, A., Linzen, T., and Marantz, A. (2014). The role of morphology in phoneme prediction: Evidence from MEG. Brain Lang. *129*, 14–23.
19. Gwilliams, L., and Marantz, A. (2015). Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words. Brain Lang. *147*, 1–13.
20. Gaston, P., and Marantz, A. (2017). The time course of contextual cohort effects in auditory processing of category-ambiguous words: MEG evidence for a single “clash” as noun or verb. Lang. Cogn. Neurosci. *33*, 402–423.
21. Singer, Y., Teramoto, Y., Willmore, B.D., Schnupp, J.W., King, A.J., and Harper, N.S. (2018). Sensory cortex is optimized for prediction of future input. eLife *7*.
22. Forseth, K.J., Hickok, G., Rollo, P.S., and Tandon, N. (2020). Language prediction mechanisms in human auditory cortex. Nat. Commun. *11*, 5240.
23. Fodor, J.A. (1985). Précis of The Modularity of Mind. Behav. Brain Sci. *8*, 1–5.
24. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., and Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science *268*, 1632–1634.
25. Tabas, A., and von Kriegstein, K. (2021). Adjudicating Between Local and Global Architectures of Predictive Processing in the Subcortical Auditory Pathway. Front. Neural Circuits *15*.
26. Jurafsky, D. (1996). A Probabilistic Model of Lexical and Syntactic Access and Disambiguation. Cogn. Sci. *20*, 137–194.
27. Norris, D., and McQueen, J.M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. Psychol. Rev. *115*, 357–395.
28. Luthra, S., Peraza-Santiago, G., Beeson, K., Saltzman, D., Crinnion, A.M., and Magnuson, J.S. (2021). Robust Lexically Mediated Compensation for Coarticulation: Christmash Time Is Here Again. Cogn. Sci. *45*.
29. Bejjanki, V.R., Clayards, M., Knill, D.C., and Aslin, R.N. (2011). Cue Integration in Categorical Tasks: Insights from Audio-Visual Speech Perception. PLOS ONE *6*, e19812.
30. Feldman, N.H., Griffiths, T.L., and Morgan, J.L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. Psychol. Rev. *116*, 752–782.
31. Chambers, C.G., Tanenhaus, M.K., and Magnuson, J.S. (2004). Actions and affordances in syntactic ambiguity resolution. J. Exp. Psychol.-Learn. Mem. Cogn. *30*, 687–696.
32. Heller, D., Parisien, C., and Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. Cognition *149*, 104–120.

33. Friston, K.J. (2010). The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. *11*, 127–138.
34. Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. *36*, 181–204.
35. McClelland, J.L., and Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychol. Rev. *88*, 375–407.
36. McClelland, J.L., and Elman, J.L. (1986). The TRACE model of speech perception. Cognit. Psychol. *18*, 1–86.
37. Magnuson, J.S., Mirman, D., Luthra, S., Strauss, T., and Harris, H.D. (2018). Interaction in Spoken Word Recognition Models: Feedback Helps. Front. Psychol. *9*, 369.
38. Swinney, D.A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. J. Verbal Learn. Verbal Behav. *18*, 645–659.
39. Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. Cognition *32*, 25–64.
40. Gaston, P., Lau, E., and Phillips, C. (2020). How does(n't) syntactic context guide auditory word recognition?
41. Altmann, G., and Steedman, M. (1988). Interaction with context during human sentence processing. Cognition *30*, 191–238.
42. Vitevitch, M.S., and Luce, P.A. (1998). When Words Compete: Levels of Processing in Perception of Spoken Words. Psychol. Sci. *9*, 325–329.
43. Vitevitch, M.S., and Luce, P.A. (1999). Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. J. Mem. Lang. *40*, 374–408.
44. Federmeier, K.D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. Psychophysiology *44*, 491–505.
45. Brodbeck, C., Hong, L.E., and Simon, J.Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. Curr. Biol. *28*, 3976-3983.e5.
46. Gillis, M., Vanthornhout, J., Simon, J.Z., Francart, T., and Brodbeck, C. (2021). Neural markers of speech comprehension: measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. bioRxiv, 2021.03.24.436758.
47. Donhauser, P.W., and Baillet, S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. Neuron *105*, 385-393.e9.
48. Ganong, W.F. (1980). Phonetic categorization in auditory word perception. J. Exp. Psychol. Hum. Percept. Perform. *6*, 110–125.
49. Leonard, M.K., Baud, M.O., Sjerps, M.J., and Chang, E.F. (2016). Perceptual restoration of masked speech in human cortex. Nat. Commun. *7*, 13619.
50. Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. Cognition *52*, 189–234.
51. Hale, J.T. (2016). Information-theoretical Complexity Metrics. Lang. Linguist. Compass *10*, 397–412.
52. Jaramillo, S., and Zador, A.M. (2011). The auditory cortex mediates the perceptual effects of acoustic temporal expectation. Nat. Neurosci. *14*, 246–251.
53. Auksztulewicz, R., Myers, N.E., Schnupp, J.W., and Nobre, A.C. (2019). Rhythmic Temporal Expectation Boosts Neural Activity by Increasing Neural Gain. J. Neurosci. *39*, 9806–9817.

54. Futrell, R., Gibson, E., and Levy, R.P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. Cogn. Sci. *44*.
55. Saffran, J.R., Aslin, R.N., and Newport, E.L. (1996). Statistical Learning by 8-Month-Old Infants. Science *274*, 1926–1928.
56. Cairns, P., Shillcock, R., Chater, N., and Levy, J. (1997). Bootstrapping Word Boundaries: A Bottom-up Corpus-Based Approach to Speech Segmentation. Cognit. Psychol. *33*, 111–153.
57. Chambers, K.E., Onishi, K.H., and Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. Cognition *87*, B69–B77.
58. Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word-recognition. Cognition *25*, 71–102.
59. Brodbeck, C., Presacco, A., and Simon, J.Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. NeuroImage *172*, 162–174.
60. Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science *343*, 1006–1010.
61. Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402.
62. Wilson, S.M., Bautista, A., and McCarron, A. (2018). Convergence of spoken and written language processing in the superior temporal sulcus. NeuroImage *171*, 62–74.
63. Lütkenhöner, B. (2003). Magnetoencephalography and its Achilles' heel. J. Physiol.-Paris *97*, 641–658.
64. McCarthy, G., and Wood, C.C. (1985). Scalp Distributions of Event-Related Potentials - an Ambiguity Associated with Analysis of Variance Models. Electroencephalogr. Clin. Neurophysiol. *61*, S226–S227.
65. Salverda, A.P., Dahan, D., and McQueen, J.M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition *90*, 51–89.
66. Beddor, P.S., McGowan, K.B., Boland, J.E., Coetzee, A.W., and Brasher, A. (2013). The time course of perception of coarticulation. J. Acoust. Soc. Am. *133*, 2350–2366.
67. Loftus, G.R., and Masson, M.E.J. (1994). Using confidence intervals in within-subject designs. Psychon. Bull. Rev. *1*, 476–490.
68. Sohoglu, E., and Davis, M.H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. eLife *9*, e58077.
69. Luke, S.G., and Christianson, K. (2016). Limits on lexical prediction during reading. Cognit. Psychol. *88*, 22–60.
70. Frisson, S., Harvey, D.R., and Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. J. Mem. Lang. *95*, 200–214.
71. Harris, Z.S. (1955). From Phoneme to Morpheme. Language *31*, 190.
72. Hitczenko, K., Mazuka, R., Elsner, M., and Feldman, N.H. (2020). When context is and isn't helpful: A corpus study of naturalistic speech. Psychon. Bull. Rev. *27*, 640–676.
73. Norris, D., McQueen, J.M., and Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. Lang. Cogn. Neurosci. *31*, 4–18.
74. DeLong, K.A., Urbach, T.P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nat. Neurosci. *8*, 1117–1121.

75. Huettig, F. (2015). Four central questions about prediction in language processing. Brain Res. *1626*, 118–135.
76. Nieuwland, M.S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. eLife *7*, e33468.
77. Pickering, M.J., and Gambi, C. (2018). Predicting while comprehending language: A theory and review. Psychol. Bull. *144*, 1002–1044.
78. Nieuwland, M.S., Barr, D.J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D.I., Ferguson, H.J., Fu, X., Heyselaar, E., Huettig, F., et al. (2020). Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. Philos. Trans. R. Soc. B Biol. Sci. *375*, 20180522.
79. Rommers, J., Meyer, A.S., Praamstra, P., and Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. Neuropsychologia *51*, 437–447.
80. Altmann, G.T.M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition *73*, 247–264.
81. Matchin, W., Brodbeck, C., Hammerly, C., and Lau, E. (2018). The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG. Hum. Brain Mapp. *40*, 663–678.
82. Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. J. Exp. Psychol. Learn. Mem. Cogn. *31*, 443–467.
83. Gazzaniga, M.S., and Sperry, R.W. (1967). Language after section of the cerebral commissures. Brain J. Neurol. *90*, 131–148.
84. Kutas, M., Hillyard, S.A., and Gazzaniga, M.S. (1988). Processing of Semantic Anomaly by Right and Left Hemispheres of Commissurotomy Patients: Evidence from Event-Related Brain Potentials. Brain *111*, 553–576.
85. Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. Trends Cogn. Sci. *9*, 512–518.
86. Federmeier, K.D., Wlotko, E.W., and Meyer, A.M. (2008). What's 'Right' in Language Comprehension: Event-Related Potentials Reveal Right Hemisphere Language Capabilities. Lang. Linguist. Compass *2*, 1–17.
87. Federmeier, K.D., and Kutas, M. (1999). Right words and left words: electrophysiological evidence for hemispheric differences in meaning processing. Cogn. Brain Res. *8*, 373–392.
88. Coulson, S., Federmeier, K.D., Van Petten, C., and Kutas, M. (2005). Right Hemisphere Sensitivity to Word- and Sentence-Level Context: Evidence From Event-Related Brain Potentials. J. Exp. Psychol. Learn. Mem. Cogn. *31*, 129–147.
89. Federmeier, K.D., Mai, H., and Kutas, M. (2005). Both sides get the point: Hemispheric sensitivities to sentential constraint. Mem. Cognit. *33*, 871–886.
90. Wlotko, E.W., and Federmeier, K.D. (2007). Finding the right word: Hemispheric asymmetries in the use of sentence context information. Neuropsychologia *45*, 3001–3014.
91. Oldfield, R.C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia *9*, 97–113.

92. Pollan, M. (2001). The Botany of Desire: A Plant's-Eye View of the World (Random House Publishing Group).
93. Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Parkkonen, L., and Hämäläinen, M.S. (2014). MNE software for processing MEG and EEG data. NeuroImage *86*, 446–460.
94. Taulu, S., and Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys. Med. Biol. *51*, 1759.
95. Bell, A.J., and Sejnowski, T.J. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Comput. *7*, 1129–1159.
96. Fischl, B. (2012). FreeSurfer. NeuroImage *62*, 774–781.
97. Hämäläinen, M.S., and Ilmoniemi, R.J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. Med. Biol. Eng. Comput. *32*, 35–42.
98. Dale, A.M., and Sereno, M.I. (1993). Improved Localizadon of Cortical Activity by Combining EEG and MEG with MRI Cortical Surface Reconstruction: A Linear Approach. J. Cogn. Neurosci. *5*, 162–176.
99. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage *31*, 968–980.
100. Brodbeck, C., Jiao, A., Hong, L.E., and Simon, J.Z. (2020). Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. PLOS Biol. *18*, e3000883.
101. Heeris, J. (2018). Gammatone Filterbank Toolkit.
102. Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2017). Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. IEEE Trans. Neural Syst. Rehabil. Eng. *25*, 402–412.
103. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In Interspeech 2017 (ISCA), pp. 498–502.
104. Brysbaert, M., and New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav Res Methods *41*, 977–90.
105. Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In Proceedings of the 6th Workshop on Statistical Machine Translation, pp. 187–197.
106. Davies, M. (2015). Corpus of Contemporary American English (COCA).
107. Lalor, E.C., Power, A.J., Reilly, R.B., and Foxe, J.J. (2009). Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. J. Neurophysiol. *102*, 349–359.
108. David, S.V., Mesgarani, N., and Shamma, S.A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. Netw. Comput. Neural Syst. *18*, 191–212.
109. Smith, S.M., and Nichols, T.E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage *44*, 83–98.

110. Greve, D.N., Van der Haegen, L., Cai, Q., Stufflebeam, S., Sabuncu, M.R., Fischl, B., and Brysbaert, M. (2013). A Surface-based Analysis of Language Lateralization and Cortical Asymmetry. J. Cogn. Neurosci. *25*, 1477–1492.

111. Bourguignon, M., Molinaro, N., and Wens, V. (2018). Contrasting functional imaging parametric maps: The mislocation problem and alternative solutions. NeuroImage *169*, 200–211.

112. Vallat, R. (2018). Pingouin: statistics in Python. J. Open Source Softw. *3*, 1026.