Neural populations in the language network differ in the size of their temporal receptive windows

Tamar I Regev[1,2*], Colton Casto[1,2*], Eghbal A Hosseini[1,2], Markus Adamek[3,4], Anthony L. Ritaccio[5], Peter Brunner[3,4,6] and Evelina Fedorenko[1,2,7]

[1] Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, Cambridge MA
[2] McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge MA
[3] National Center for Adaptive Neurotechnologies, Albany NY
[4] Washington University School of Medicine, St. Loius MO
[5] Department of Neurology, Mayo Clinic, Jacksonville FL
[6] Albany Medical Center, Albany NY
[7] Speech and Hearing Bioscience and Technology (SHBT) Program, Harvard University, Boston MA

* Equal contribution

Correspondence should be sent to: tamarr@mit.edu; ccasto@mit.edu; evelina9@mit.edu

Acknowledgements

We would like to acknowledge the participants for agreeing to take part in our study, as well as Nancy Kanwisher, former and current EvLab members, especially Cory Shain and Anya Ivanova, and the audience at the Neurobiology of Language conference (2022, Philadelphia) for helpful discussions and comments on the analyses and manuscript. TIR was supported by the Zuckerman-CHE STEM Leadership Program and by the Poitras Center for Psychiatric Disorders Research. EF was supported by NIH awards R01-DC016607, R01-DC016950, and U01-NS121471 and research funds from the McGovern Institute for Brain Research, Brain and Cognitive Sciences department, and the Simons Center for the Social Brain.

Abstract

Language comprehension is fast and seemingly effortless. However, in spite of long knowing what brain regions enable this feat, our knowledge of the precise neural computations that these frontotemporal regions implement remains limited. One highly controversial question is whether there exist functional differences among the neural populations that comprise the language network. Leveraging the high spatiotemporal resolution of intracranial recordings, we clustered the timecourses of responses to sentences and linguistically degraded conditions and discovered three response profiles that robustly differ in their temporal dynamics. Computational modeling of these profiles suggested that they reflect different temporal receptive windows (TRWs), with average TRWs of 1, 4, and 6 words. The electrodes exhibiting these profiles were interleaved across the language network, further suggesting that all language regions have direct access to distinct, multi-scale representations of linguistic input—a property that may be critical for the efficiency and robustness of language processing.

Introduction

Language processing engages a network of brain regions that reside in the temporal and frontal lobes and are typically left-lateralized (e.g., Fedorenko et al., 2010; Pallier et al., 2011). These brain regions respond robustly to linguistic stimuli across presentation modalities (Fedorenko et al., 2010; Vagharchakian et al., 2012; Regev et al., 2013; Scott et al., 2017), tasks (Fedorenko et al., 2010; Cheung et al., 2020; Diachek, Blank, Siegelman et al., 2020), and languages (Malik-Moraleda, Ayyash et al. 2022). This language-responsive network is highly selective for language, showing little or no response to diverse non-linguistic stimuli (e.g., Fedorenko et al., 2011; Monti et al., 2012; Deen et al., 2015; Ivanova et al., 2020, 2021; Liu et al., 2020; Chen et al., 2023). However, the precise computations and neuronal dynamics that underlie language comprehension remain debated.

Based on neuroimaging and aphasia evidence, some have argued for dissociations among different aspects of language, including phonological/word-form processing (e.g., Okada and Hickok, 2006; Graves et al., 2008; DeWitt and Rauschecker, 2012), the processing of word meanings (e.g., Price et al., 1997; Rodd et al., 2005; Mesulam et al., 2013), and syntactic/combinatorial processing (e.g., Friederici, 2002, 2011; Hagoort, 2005; Grodzinsky and Santi, 2008; Matchin and Hickok, 2020). However, other studies have reported distributed sensitivity to these aspects of language across the language network (Fedorenko et al., 2010, 2020; Bautista and Wilson, 2016; Blank et al., 2016; Huth et al., 2016; Shain, Blank et al., 2020; Regev et al., 2021). Some of the challenges in arriving at a clear answer may have to do with the limitations of the dominant methodologies available for studying language processing. For example, fMRI examines neural responses in voxels that are a few cubic millimeters in size. Each voxel contains a million or more individual neurons, which may differ functionally. If different linguistic computations are implemented in distinct neural populations that are distributed and interleaved across the language cortex, such dissociations may be difficult to detect with fMRI. Further, fMRI measures neural activity averaged across time (typically, every ~2 seconds), which may obscure linguistic computations that happen on a faster timescale (**Figure 1**).

In recent years, invasive recordings of human neural activity (e.g., Mukamel and Fried, 2011), including electrocorticography (ECoG) and stereo electroencephalography (sEEG), have become increasingly available to language neuroscience researchers, as patients undergoing presurgical evaluation (usually for intractable epilepsy) agree to perform linguistic tasks while implanted with intracranial electrodes. These data have high spatial resolution (standard macro-electrodes record activity of relatively small populations of neurons) and high temporal resolution (millisecond-level), allowing the tracking of neural dynamics across both space and time. As a result, intracranial recordings have the potential to uncover both a) nearby electrodes that show distinct functional profiles (whose responses would be averaged in fMRI) (**Figure 1A**), and b) electrodes that show distinct activity patterns over time at the scale inaccessible to fMRI (**Figure 1B**).

Several previous studies have probed intracranial neural responses during language comprehension (e.g., Fedorenko et al., 2016; Nelson et al., 2017). For example, Fedorenko et al.

(2016) reported sensitivity in language-responsive electrodes to both word meanings and combinatorial processing, in line with fMRI findings (e.g., Fedorenko et al., 2010; Bedny et al., 2011). They also reported a temporal profile where neural activity gradually increases (builds up) across the sentence (replicated by Nelson et al., 2017), which they interpreted as reflecting the construction of a sentence meaning. However, only a subset of the language-responsive electrodes showed this profile, leaving open the questions of what other response profiles may exist within the language network and what computations those profiles may be associated with.

Here, we report a detailed investigation of neural responses during language processing. We focus on language-responsive electrodes (as in Fedorenko et al., 2016), which respond reliably more strongly to sentences compared to sequences of pseudowords. To foreshadow our findings, we report three response types that exhibit distinct temporal dynamics and vary in their response magnitudes to different linguistic conditions and in their degree of locking to the stimulus. We argue that these response types relate to the timescales of information integration in the language system (e.g., Lerner et al., 2011; Blank and Fedorenko, 2020), and use a simplified model of neural responses to estimate the temporal receptive window for each response type.

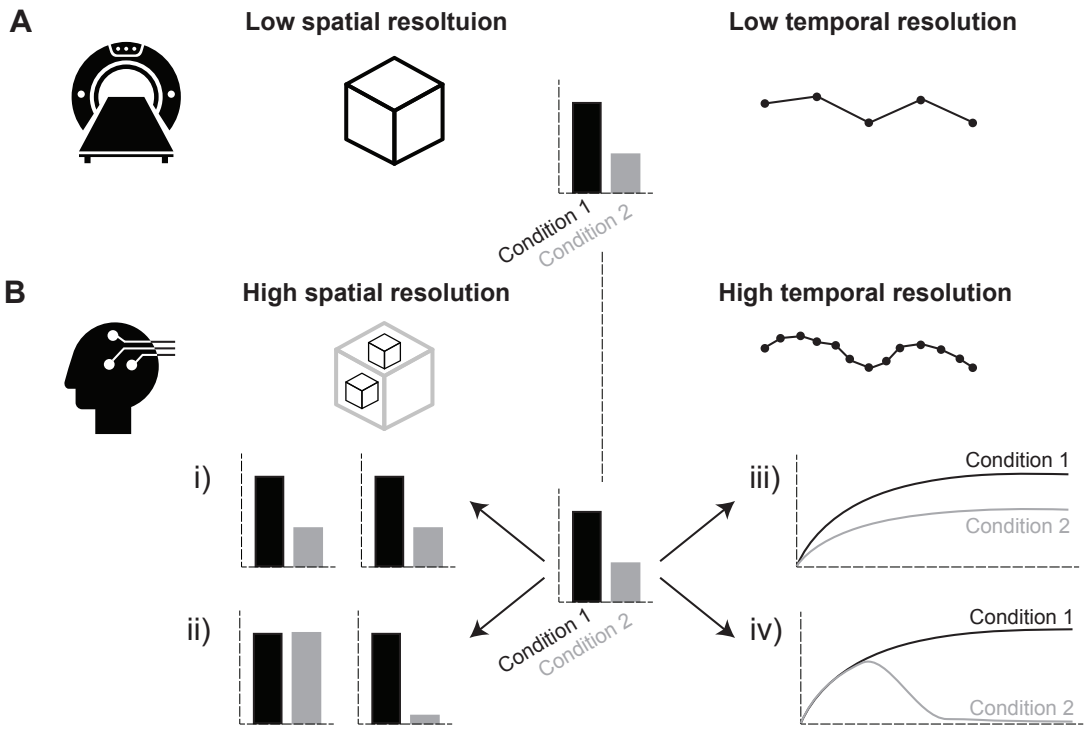

Figure 1 – **Low spatial and temporal resolution may obscure details of neural computations**. **A)** Hypothetical effect sizes for 2 experimental conditions as measured using fMRI, with relatively low spatial and temporal resolution. **B)** Illustration of several possible outcomes of the same hypothetical experimental conditions when measured with intracranial EEG (ECoG/sEEG), which has both high spatial and high temporal resolution. The high spatial resolution may reveal the same effect size in smaller neural units (i) or, alternatively, it may reveal distinct functional profiles (including those that differ qualitatively from the average) (ii). The high temporal resolution may reveal distinct temporal dynamics as well (iii, iv).

Results

We used intracranial recordings from patients with intractable epilepsy to investigate neural responses during language comprehension. Participants in Dataset 1 were presented with four types of linguistic stimuli that have been traditionally used to tease apart neural responses to word meanings and syntactic structure (Fedorenko et al., 2010, 2012, 2016; for earlier uses of this paradigm, see Mazoyer et al., 1993; Friederici et al., 2000; Humphries et al., 2001; Vandenberghe et al., 2002): sentences (S), lists of unconnected words (W), Jabberwocky sentences (J), and lists of nonwords (N) (**Figure 2A,B, Methods,** all stimuli are available at osf.io/xfbr8/). In each trial, 8 words or nonwords were presented on a screen serially and participants were asked to silently read them. To maintain alertness, after each trial participants judged whether a probe word/nonword had appeared in that trial. See **Methods** for further details of stimulus presentation. In Dataset 2, just two of these conditions were used: sentences and lists of nonwords.

We asked three research questions: 1) Does the language network contain reliably distinct response profiles? If so - 2) What do these profiles reflect? And finally - 3) Do electrodes with different response profiles tend to be located in particular regions of the language network? We used Dataset 1 for initial evaluation of these questions because this dataset contained a richer set of experimental conditions. We then used Dataset 2 as an attempt to replicate the findings despite the more compact experimental paradigm.

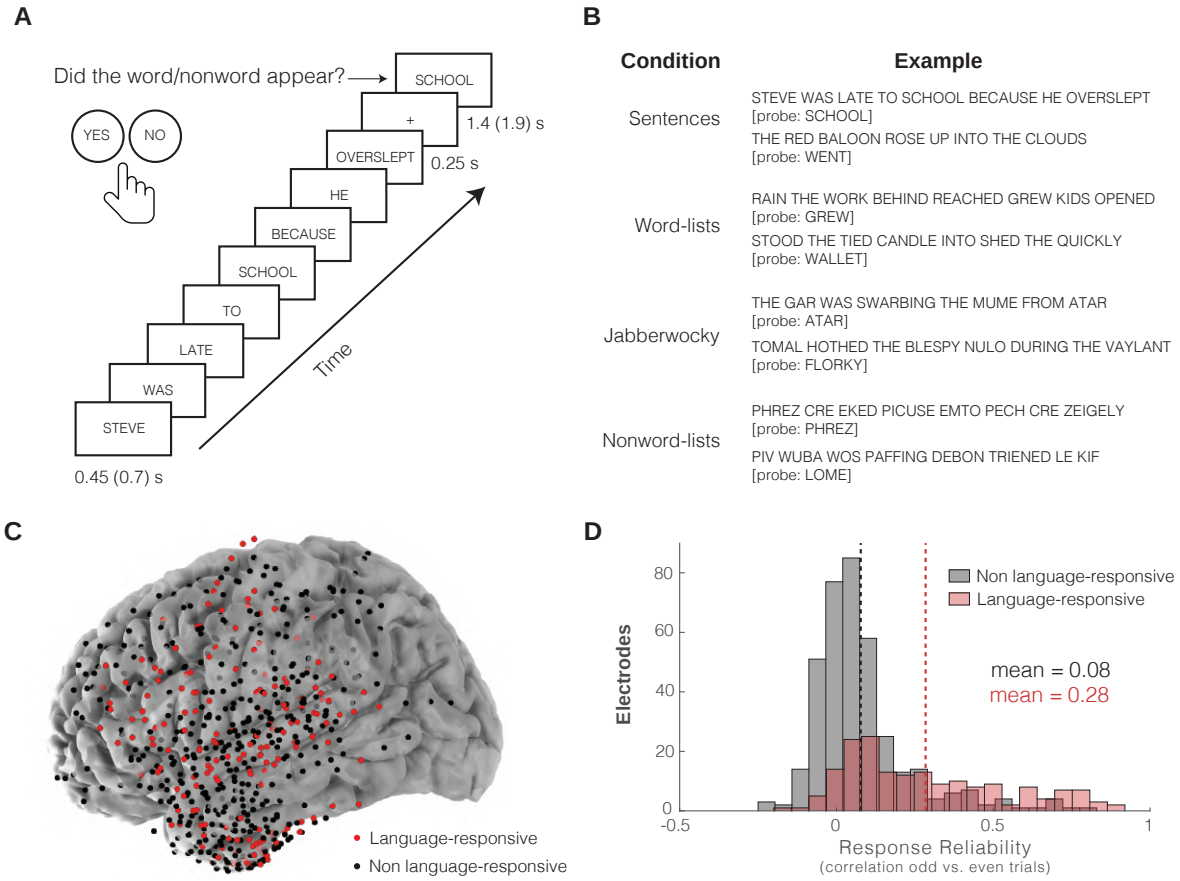

Figure 2 – Experimental procedure and the distribution of the implanted electrodes for Dataset 1. **A)** A sample trial from the sentences condition. **B)** Two sample items from each of the four experimental conditions. **C)** The locations of language-responsive (n=177, red; see Language-Responsive Electrode Selection) and non language-responsive (n=373, black) electrodes across the six participants in Dataset 1. Electrodes were implanted almost exclusively in the left hemisphere for Dataset 1 and concentrated in the temporal and frontal lobes, with language-responsive electrodes found across the cortex. **D)** Response reliability across odd and even trials (based on a correlation of mean responses) for language-responsive and non language-responsive electrodes. Language-responsive electrodes exhibit more reliable responses to linguistic stimuli than non language-responsive electrodes.

1. Language-responsive electrodes exhibit reliably distinct response profiles.

We clustered the gamma-band neural response patterns of language-responsive electrodes from Dataset 1 (6 participants, 177 language-responsive electrodes; **Figure 2C**, Methods, **Table 1**) to sentences (S), word-lists (W), Jabberwocky sentences (J) and nonword-lists (N) (**Figure 2A, B**). The k-medoids clustering algorithm (see **Figure S1** for evidence that similar results emerge with a k-means clustering algorithm), combined with the "elbow" method (Methods), suggested that three clusters (k=3) optimally explained the data (**Figure 3A**). Although we combined the electrodes from all 6 participants for clustering, electrodes that belong to Cluster 1 (92 total electrodes) and Cluster 2 (67 total electrodes) were evident in every participant individually, and

electrodes that belong to Cluster 3 (18 total electrodes) were evident in 4 of the 6 participants (**Figure 3B**). Furthermore, electrodes that exhibit more reliable responses to linguistic stimuli (**Figure 2D**, Methods) show clearer separation among the clusters in a low-dimensional space (**Figure 3C**), which suggests that these clusters appropriately characterize the underlying structure of language-selective responses.

Inspection of the average timecourse by cluster (**Figure 3E**) revealed three distinct response profiles (see **Figure 3D** for best representative electrodes—medoids chosen by the k-medoids algorithm—from each cluster). Cluster 1 (n=92 electrodes) was characterized by a relatively slow increase (build-up) of neural activity across the 8 words in the S condition (a pattern similar to the one reported by Fedorenko et al., 2016 and Nelson et al., 2017; see Discussion), and much lower activity for the W, J, and N conditions, with no difference between the J and N conditions (**Figure 3F**). Cluster 2 (n=67 electrodes) displayed a quicker build-up of neural activity in the S condition that plateaued approximately 3 words into the sentence, a quick build-up of activity in the W condition that began to decay after the third word, and a similar response to the J and N conditions as to the W condition with an overall lower magnitude. Cluster 2 also exhibited 'locking' of the neural activity to the onsets of individual words in the S condition. Finally, Cluster 3 (n=18 electrodes) showed no build-up of activity, and instead was characterized by a high degree of locking to the onset of each word or nonword in all conditions. Additionally, the response magnitudes of Cluster 3 were more similar across conditions compared to the other two clusters, although the S>W>J>N pattern was still present (**Figure 3F**).

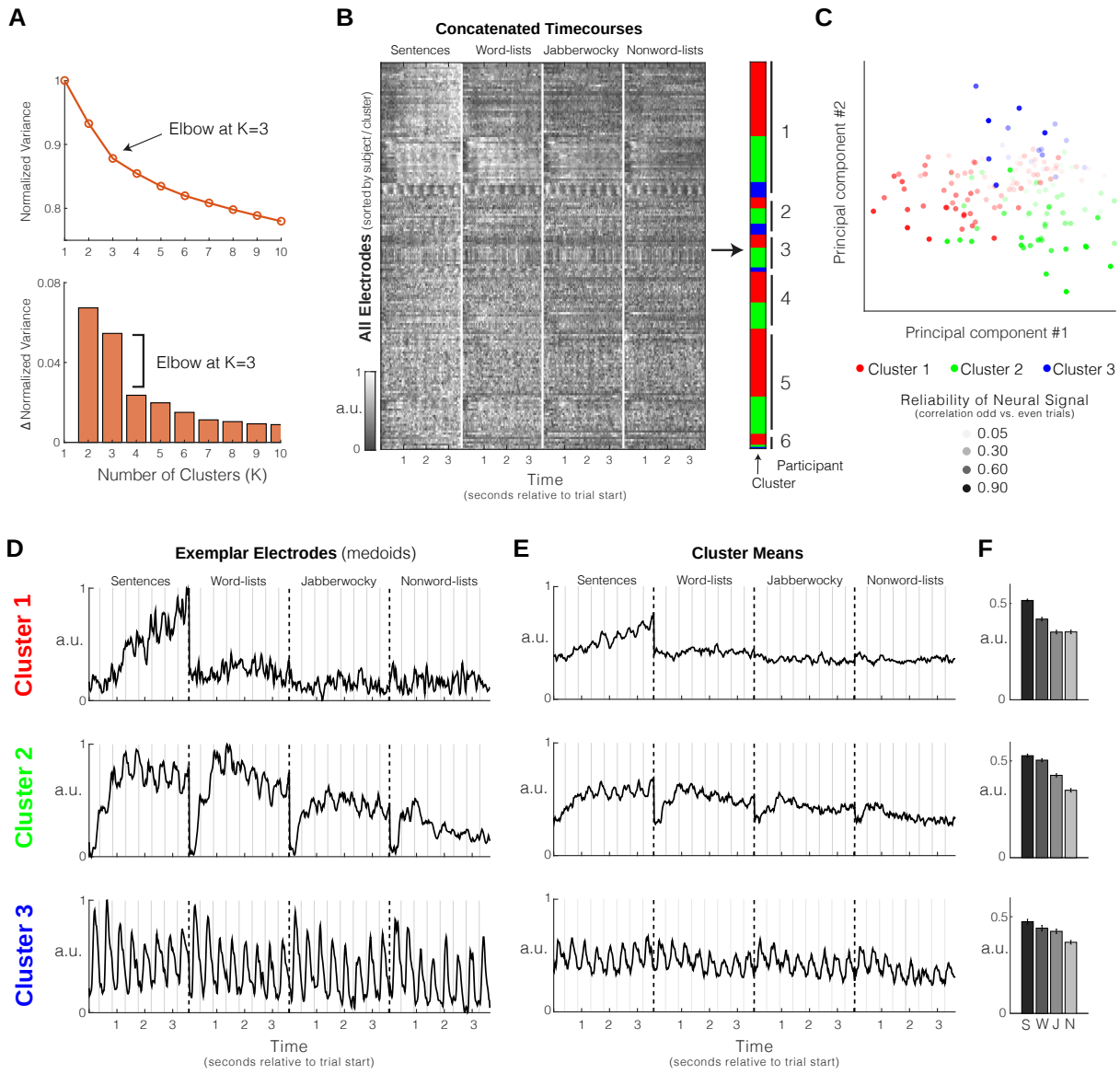

Figure 3 – Dataset 1 k-medoids clustering results. A) Search for optimal k using the "elbow method". Top: variance (sum of the distances of all electrodes to their assigned cluster center) normalized by the variance when k=1 as a function of k (normalized variance (NV)). Bottom: change in NV as a function of k (NV(k+1) – NV(k)). After k=3, there is a large drop in the change in variance, suggesting that 3 clusters appropriately describe this dataset. **B)** Clustering mean electrode responses (concatenated across the four experimental conditions: sentences (S), word-lists (W), jabberwocky (J), nonword-lists (N)) using k-medoids (k=3) with a correlation-based distance (Methods). Shading of the data matrix reflects normalized high-gamma power (70-150Hz). Clusters 1 and 2 (red and green, respectively) are present in all six participants, and Cluster 3 (blue) is present in four of six participants. **C)** Electrode responses visualized on their first two principal components, colored by cluster and shaded by the reliability of the neural signal as estimated by correlating responses to odd and even trials (**Figure 2D**). **D)** Timecourses of best representative electrodes ('medoids') selected by the algorithm from each of the three clusters. The timecourses reflect normalized high-gamma (70-150Hz) power averaged over all trials of a given condition. **E)** Timecourses averaged across all electrodes in each cluster. Clusters exhibit distinct response profiles (quantitatively described in **Figure 5**). **F)** Mean condition responses by cluster. Error bars reflect standard error. After averaging across time, response profiles are not as distinct by cluster (especially for Clusters 2 and 3), underscoring the importance of temporal information in elucidating this grouping of electrodes.

We then evaluated how stable these clusters were across trials and how robust they were to data loss. We found that the same clusters emerged when using only half of the trials in Dataset 1 (either odd- or even-numbered trials): clusters derived from half of the data were significantly more similar to the clusters derived from the full dataset than would be expected by chance (ps<0.001 for all 3 clusters for each half of the data, evaluated against clustering solutions from permuted data, Methods, **Figure S2A**). The clusters were also robust to the number of electrodes used: clustering solutions derived from only a subset of the language-responsive electrodes (down to ~27%, ~32%, and ~69% of electrodes for clusters 1, 2, and 3, respectively) were significantly more similar to the clusters derived from all the electrodes than would be expected by chance (ps<0.05, Methods, **Figure S2B**). Although this result suggests that the reported clustering solution is robust to data loss, it also demonstrates that, relative to the other clusters, Cluster 3 is more strongly driven by individual electrodes (which is to be expected given that Cluster 3 has the fewest electrodes assigned to it, **Figure 3B**). In sum, the three response profiles discovered in Dataset 1 were stable across trials and robust to data loss.

2. Response profiles reflect different sizes of temporal receptive windows

The temporal dynamics of the neural responses across clusters suggested that the observed differences in the response profiles may reflect different 'temporal receptive windows' (TRWs). A TRW is a temporal equivalent of spatial receptive fields that corresponds to the amount of the preceding temporal context that affects the processing of the current input (e.g., Hasson et al., 2008). More specifically, two features of the temporal dynamics of the neural responses pointed to distinct TRWs. The first is the difference in relative response magnitudes to the linguistic experimental conditions. A larger difference between sentences and word-lists (Cluster 1) implies a longer TRW as the response is strongly modulated by information that spans multiple words (i.e., combinatorial information present in sentences but not word-lists). Similarly, a smaller difference between sentences and word-lists and between word-lists and nonword-lists (Clusters 2 and 3) implies a shorter TRW because the response is only modulated slightly by the addition of information that spans multiple words and single words, respectively. The second feature is the difference in the degree of 'locking' to individual word onsets, which manifests as oscillations at the rate of stimulus presentation, where stronger locking (Cluster 3) implies a shorter TRW given the sensitivity to information at the word level (or below). We therefore hypothesized that Cluster 1 was dominated by neural populations with the longest TRWs, Cluster 2 by neural populations with shorter TRWs, and Cluster 3 by neural populations with the shortest TRWs.

To evaluate whether the response profiles that we uncovered indeed correspond to varying TRW sizes, as well as to quantitatively estimate the sizes of the TRWs, we turned to modelling. In particular, we simulated neural responses to the sentence condition by convolving a simplified stimulus train with a range of kernels that represent the size of the TRW (σ**, Figure 4A**, Methods). The kernels were constructed from gaussian curves with a standard deviation of $\sigma/2$ truncated at +/- 1 standard deviation (capturing 2/3 of the area under the gaussian, **Figure 4a**, Methods), and σ varied from one third of a word to 8 words (the length of our stimuli), in increments of $1/27^{th}$ of a word (or (60Hz)—the highest resolution we were able to evaluate given the sampling

rate). The simulated neural responses exhibit a striking qualitative similarity to the observed response patterns (**Figure 4A**). We then computed—for every electrode—a correlation between each simulated response and the observed response. The TRW of an electrode was defined as the σ that yielded the highest correlation (Methods).

The estimated TRW sizes showed a clear pattern of Cluster 1>2>3; the average σ values per cluster were ~6, ~4, and ~1 words for Clusters 1, 2, and 3, respectively (p-values comparing all pairs of clusters <0.0001, LME, Methods, **Figure 4B, C, Table S5**). Furthermore, in order to test whether the estimated values of σ depended on the stimulus presentation rate, which varied across participants, we calculated the average σ per cluster separately for the participants that chose a faster (n=3) vs. a slower (n=3) presentation rate. The estimated values of σ in number of words (as reported above) were invariant to the presentation rate (**Figure S6**). This invariance suggests that the TRW of language-responsive electrodes is information-, not time-, dependent.

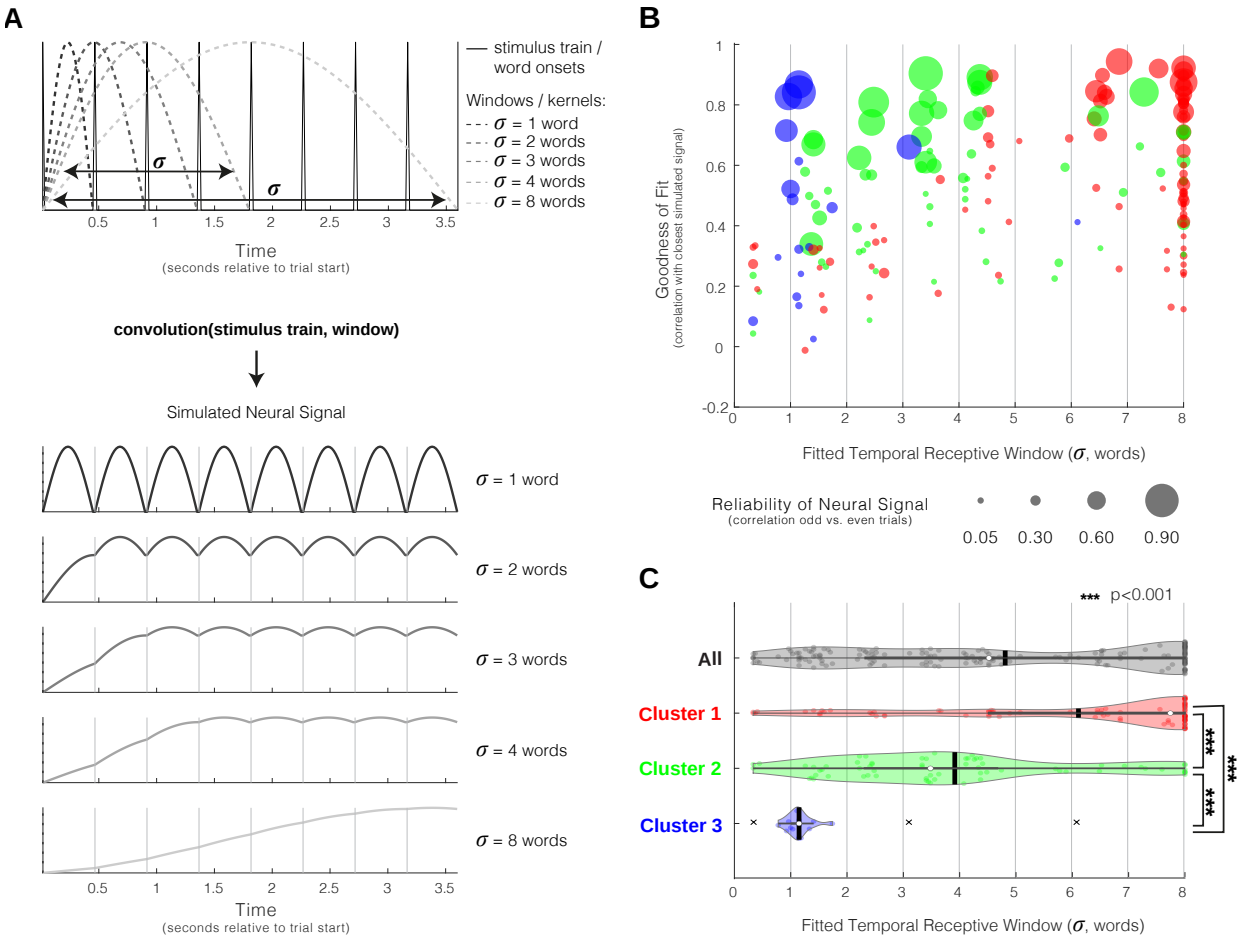

Figure 4 – Estimating the size of the temporal receptive window (TRW) of different electrodes. **A)** A model that simulates neural responses to the sentence condition as a convolution of a simplified stimulus train and gaussian kernels with varying widths. Top: Simplified stimulus train where peaks indicate a word/nonword onset, and sample kernels of varying temporal receptive window sizes (σ). The kernels were constructed from gaussian curves with a standard deviation of $\sigma/2$ truncated at +/- 1 standard deviation (capturing 2/3 of the area under the gaussian, Methods) and normalized to a minimum of 0 and a maximum of 1. Bottom: The resulting

simulated neural signals for sample kernel widths, normalized to a minimum of 0 and a maximum of 1. **B)** Best TRW fit for all electrodes colored by cluster and sized by the reliability of the neural signal as estimated by correlating responses to odd and even trials (**Figure 2D**). The goodness of fit, or correlation between the simulated and observed neural signal (sentence condition only), is shown on the y-axis. **C)** Estimated TRW sizes across all electrodes (grey) and per cluster (red, green, and blue). Black vertical lines correspond to the mean window size and the white dots correspond to the median. "x" marks indicate outliers (more than 1.5 interquartile ranges above the upper quartile or less than 1.5 interquartile ranges below the lower quartile). Significance values are calculated using a linear mixed-effects model (ps<0.001, LME, Methods). Together, B and C show that our model explains the observed neural signals well and that the clusters vary in the size of their TRWs, from a relatively long TRW (Cluster 1) to a relatively short one (Cluster 3).

To further quantify the apparent differences in the response profiles, we performed two additional analyses (Methods). First, we examined how quickly the S condition diverges from the W condition (**Figure 5A**), and how quickly the W condition diverges from the N condition (**Figure 5B**), using a binary linear classifier trained for each cluster separately using incrementally more of the timecourse (from one time bin to the entire timecourse, Methods). The average classification performance (across 20 unique instantiations of a 10-fold cross-validated binary classifier) revealed that electrodes in Cluster 1 reliably distinguish both S from W and W from N already in the earliest time bins. In contrast, electrodes in Cluster 2 reliably distinguish S from W and W from N starting at word positions 2-3 and onward, and electrodes in Cluster 3 do not reliably distinguish W from N in any continuous stretches of time and only distinguish S from W with access to nearly the entire timecourse.

Second, we examined how strongly the neural signal exhibits 'locking' to individual word/nonword onsets by correlating the observed responses with a fitted sinusoidal stimulus train (Methods). This analysis revealed that—consistent with visual examination—electrodes in Cluster 3 show the strongest degree of stimulus locking, followed by electrodes in Cluster 2, with electrodes in Cluster 1 showing the weakest stimulus-related time-locking (**Figure 5C, Table S1).** This difference in the degree of stimulus locking was present across conditions, although the analysis additionally revealed a by-condition trend (**Figure 5C**) that did not reach significance (**Table S1**), with strongest stimulus locking in the S condition and weakest stimulus locking in the N condition for all three clusters (with no reliable cluster by condition interaction). These qualitative between-condition differences could be due to generally greater engagement with more language-like stimuli.

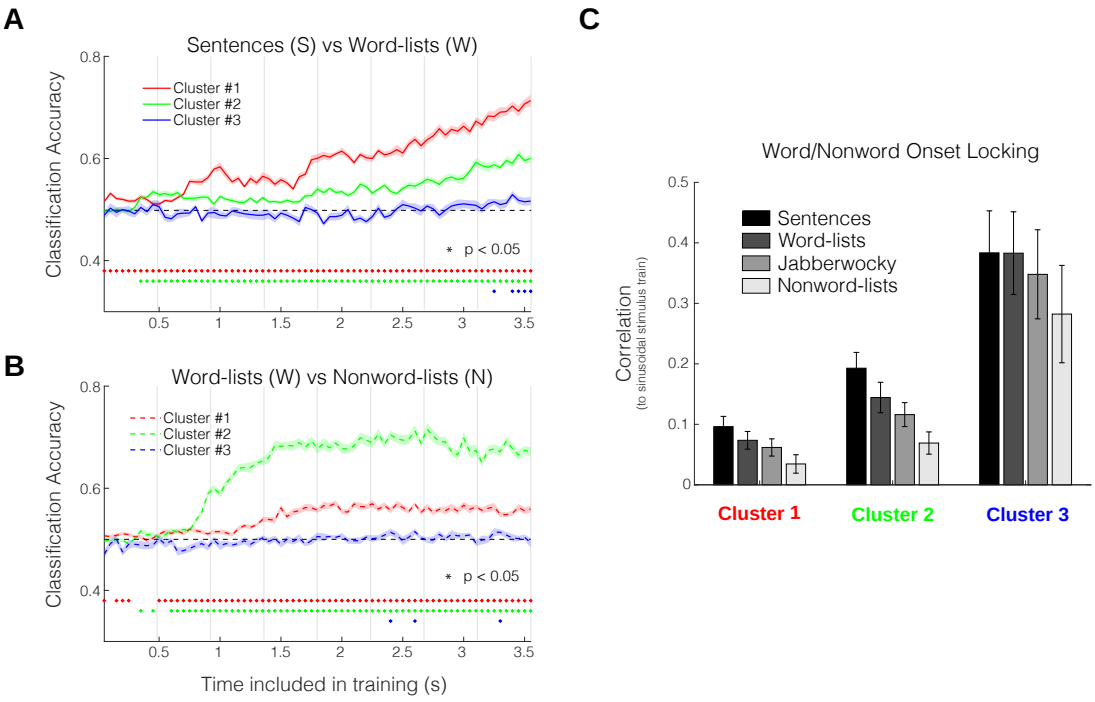

Figure 5 – Quantitative characterization of the three clusters in Dataset 1. A and B) Classifier performance by cluster as a function of the amount of timecourse included in training (Methods). Classifiers (n=20) were trained to discriminate sentence (S) and word-list (W) conditions (**A**) and word-list (W) and nonword-list (N) conditions (**B**). Significance stars (colored by cluster) reflect discriminability of conditions above chance level (0.50) using a one-tailed t-test (ps<0.05). Shaded regions reflect standard error. The responses to sentences and word-lists are most discriminable for electrodes in Cluster 1, whereas the responses to word-lists and nonword-lists are most discriminable for electrodes in Cluster 2. Responses to sentences and word-lists as well as nonword-lists cannot be discriminated for electrodes in Cluster 3 (until the classifier has access to the entire timecourse in the case of discriminating sentences and word-lists). **C)** Correlation of fitted stimulus train with timecourse of electrodes by cluster and by condition (Methods). Error bars reflect standard error. Electrodes in Cluster 3 are the most locked to word/nonword presentation while electrodes in Cluster 1 are the least locked to word/nonword presentation.

These differences among the clusters in their ability to discriminate between conditions and their degree of stimulus locking provide additional evidence in support of the idea that the response profiles differ in the size of the TRW.

3. Clusters 1 and 2 are distributed across the language network, whereas cluster 3 exhibits a posterior bias.

We tested for differences in the anatomical distribution of the electrodes that belong to the 3 clusters in Dataset 1. We excluded from this analysis right-hemisphere (RH) electrodes because only 4 RH electrodes passed the language selectivity criterion (S>N). We therefore focused on the y (posterior-anterior) and z (inferior-superior) directions within the left hemisphere.

Electrodes from both Cluster 1 and 2 were distributed across the temporal and frontal language regions (**Figure 6**). When examining all electrodes together or only the frontal electrodes, the MNI-coordinates of Clusters 1 and 2 did not significantly differ in either of the two tested directions (ps>0.05, LME, **Figure 6C, D**, Methods**, Table S2**). However, when examining only the electrodes located in the temporal lobe, electrodes from Cluster 1 were more inferior relative to Cluster 2 (p<0.001, LME, Methods, **Figure 6E, Table S2**), reflecting a large proportion of electrodes belonging to Cluster 1 on the ventral temporal surface.

Electrodes from Cluster 3 were significantly more posterior than those in Clusters 1 and 2 (Cluster 3 vs. 1: p<0.001, LME, Methods, **Figure 6C, Table S2**; Cluster 3 vs. 2: p<0.01, LME, Methods, **Figure 6C, Table S2**). This trend was also evident when examining the temporal and frontal electrodes separately, but the difference only reached significance for the temporal electrodes (**Figure 6D, E**).

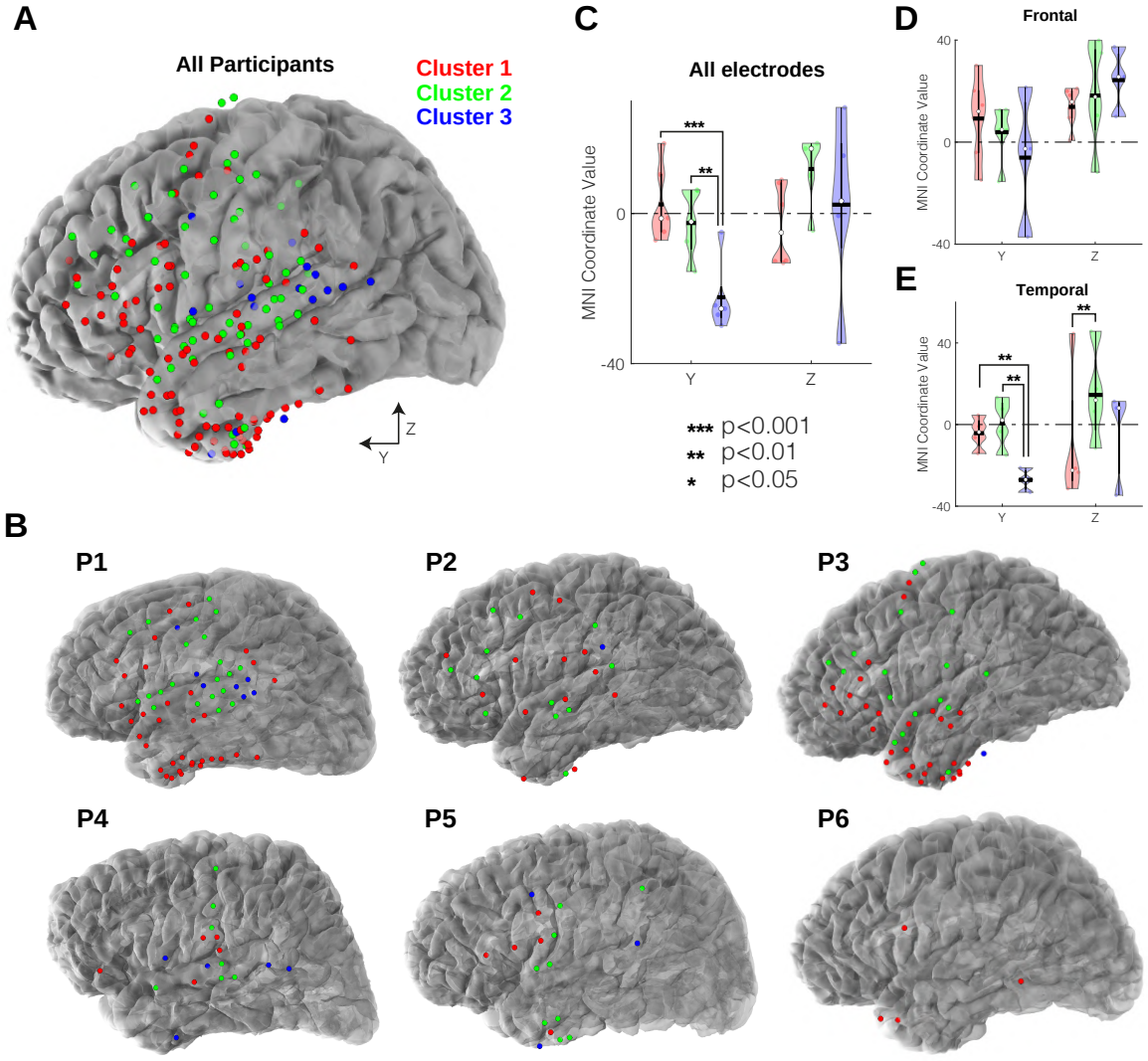

Figure 6 – Anatomical distribution of the clusters in Dataset 1. A) Anatomical distribution of language-responsive electrodes in Dataset 1 across all participants in MNI space, colored by cluster. **B)** Anatomical distribution of language-responsive electrodes in participant-specific space. **C-E)** Violin plots of MNI coordinate values for the 3 clusters, where plotted points represent the mean of all coordinate values for a given participant and cluster. The mean is plotted with a black horizontal line, and the median is shown with a white circle. Significance values are computed using a linear mixed-effects model (LME, Methods). Cluster 3 exhibits a posterior bias (more negative Y coordinate) relative to Cluster 1 and 2 when modeled using all language electrodes (**C**)**.** When electrodes are modeled separately by lobe, Cluster 1 shows a significant inferior bias (more negative Z coordinate) relative to Cluster 2 in the temporal lobe (**E**).

4. Clusters 1 and 3 replicate in Dataset 2 and cluster 2 partly replicates.

We then asked whether the same clusters would emerge in a second, independent dataset with new participants and linguistic materials (Dataset 2; 16 participants; 362 language-responsive

electrodes; mostly depth electrodes, **Figure 7A**). Participants in Dataset 2 only saw two of the four conditions presented to participants in Dataset 1 (sentences (S) and nonword-lists (N), but not word-lists (W) and Jabberwocky sentences (J)); therefore, we first re-clustered the electrodes from Dataset 1 using only the responses to the S and N conditions.

The optimal number of clusters when only the S and N conditions were used from Dataset 1 was again k=3, and the cluster centers exhibited striking qualitative similarity to those of the clusters derived using the data from all four conditions (**Figure S3**). In line with this similarity, ~80% of electrodes were assigned to the same cluster across these two clustering solutions. This result suggests that it is possible to recover the 3 response profiles from only responses to sentences and nonword-lists. However, the analysis where varying subsets of electrodes were removed (**Figure S3G**) revealed that Cluster 2 was less robust than Clusters 1 and 3 to electrode loss when only the S and N conditions were used (compare the green curve in **Figure S2B** to the green curve in **Figure S3G**). This pattern suggests that responses to the word-list (W) and Jabberwocky sentence (J) conditions may be especially important for differentiating Cluster 2 from the other response profiles.

We next clustered the electrodes in Dataset 2 using the same approach as for Dataset 1 (k-medoids algorithm, correlation-based distance). The optimal number of clusters in Dataset 2 was k=2 based on the elbow method (**Figure S4A**), and the resulting clusters were reliably similar to Clusters 1 and 3 from Dataset 1 ($ps<0.001$ for both clusters, **Figure S4C**). However, to test whether Cluster 2 could emerge from Dataset 2, we also clustered Dataset 2 enforcing k=3. When the electrodes in Dataset 2 were clustered into three sets, the same two cluster centers as in the case of k=2 were again apparent and showed reliable similarity to Clusters 1 and 3 in Dataset 1 ($p<0.001$ and $p<0.05$, respectively, **Figure 7G, I**). The third cluster qualitatively resembled Cluster 2 from Dataset 1 (an initial build-up of activity followed by a plateau (S condition) or decay (N condition); **Figure 7G**), but the resemblance was not statistically reliable. This result may be, in part, attributable to i) the lower data quality of Dataset 2 compared to Dataset 1 (compare **Figure 7B** vs. **7C**); and ii) the greater sparsity of coverage due to the prevalence of depth electrodes. Additionally, see **Figure S5** (and **Table S3** and **S4**) for an analysis of the anatomical trends in Dataset 2 which showed weak, but not reliable, differences between Clusters 1 and 3.

We then estimated the temporal receptive window (TRW) size (as in Section 2 above) for each electrode in Dataset 2. Clusters 1 and 3 in Dataset 2 (which replicated the findings from Dataset 1 in terms of their profiles) are best described by TRWs of ~4.5 and ~1, respectively (**Figure S7A, B**), similar to Dataset 1.

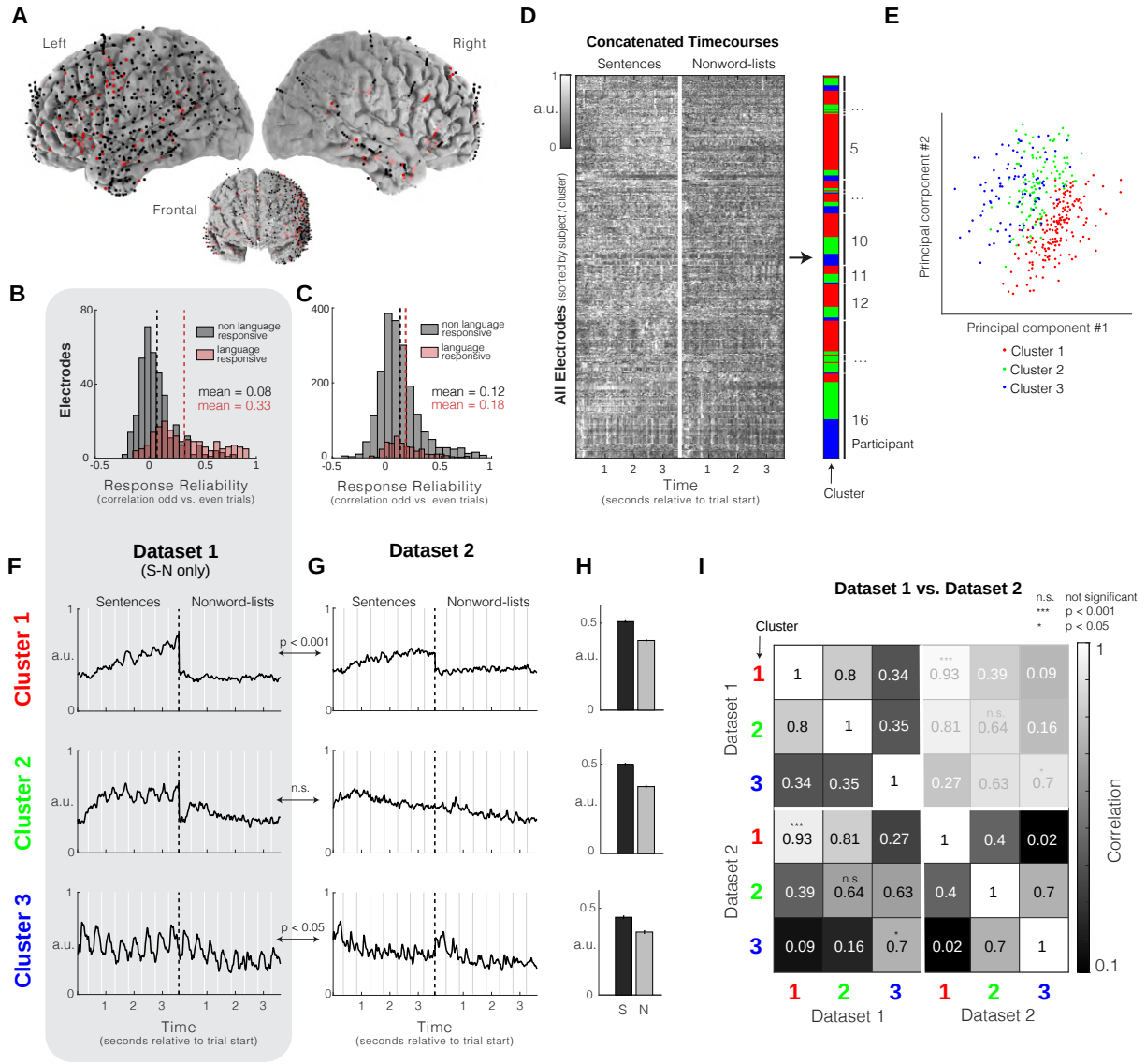

Figure 7 – Dataset 2 k-medoids clustering results (k=3). A) The locations of language-responsive (n=362, red; Methods) and non-language-responsive (n=2,017, black) electrodes across the sixteen participants in Dataset 2 (both surface and depth electrode were implanted). Language-responsive electrodes were found across the cortex, in both the left and right hemispheres (**Table 2**). **B and C)** Response reliability across odd and even trials (based on a correlation of mean responses) for language-responsive and non language-responsive electrodes (as in **Figure 2D**). Language-responsive electrodes exhibit more reliable responses to linguistic stimuli than non language-responsive electrodes for both Dataset 1 (S+N conditions only, **B**) and Dataset 2 (**C**), however, the responses of language electrodes are less reliable in Dataset 2 than Dataset 1. **D)** Clustering mean electrode responses (S+N) in Dataset 2 using k-medoids (k=3) with a correlation-based distance. Shading of the data matrix reflects normalized high-gamma power (70-150Hz). **E)** Electrodes visualized on their first two principal components, colored by cluster. **F and G)** Average timecourse by cluster from Dataset 1 when using only S and N conditions (**F**; see **Figure S2**) and from Dataset 2 (**G**). **H)** Mean condition responses by cluster in Dataset 2. Error bars reflect standard error. Again, after averaging across time, response profiles are not as distinct by cluster, underscoring the importance of temporal information in elucidating this grouping of electrodes. **I)** Evaluation of clusters from Dataset 1 (clustering with S and N conditions only) against clusters from Dataset 2. Clusters 1 and 3 from Dataset 1 replicate to Dataset 2 ($p<0.001$ and $p<0.05$, respectively; evaluated against clustering solutions when trials are shuffled; Methods), and although Cluster 2 demonstrates some qualitative similarity across the two datasets, this similarity is not statistically reliable.

Discussion

The nature of the neural computations that support our ability to extract meaning from linguistic input remains an important open question in the field of language research. Here, we leveraged the high temporal and spatial resolution of human intracranial recordings to probe the fine temporal dynamics and the spatial distribution of language-responsive electrodes. We uncovered three temporal profiles of response during the processing of sentences and linguistically degraded conditions like lists of words or nonwords. We suggest that these profiles reflect different sizes of the temporal receptive window (TRW)—the amount of temporal context that affects the processing of the current input. Further, we found that electrodes that exhibit these response profiles do not co-localize to the same parts of the language network, instead manifesting in a 'salt and pepper' pattern across both frontal and temporal cortex. Below, we contextualize these results with respect to prior empirical work and discuss their implications for our understanding of human language processing.

Three temporal profiles characterize language-responsive electrodes

In the present study, we used a clustering approach to group intracranial electrodes by their responses to four types of language stimuli: sentences (S), lists of unconnected words (W), Jabberwocky sentences (where content words are replaced with pronounceable nonwords; J), and lists of nonwords (N). We uncovered three response profiles. One set of electrodes (~52% of the language-responsive electrodes) showed a slow increase (build-up) of neural activity across the words in the sentence, and much lower activity for the three linguistically degraded conditions. Another set of electrodes (~38% of the language-responsive electrodes) showed a faster build-up across the words in the sentence, plateauing at ~3 words into the sentence and exhibiting some degree of 'locking' to individual word/nonword onsets; the response during the first three words resembled that for the word-list condition, but then the response decayed; the remaining two conditions showed overall lower responses but a similar shape as the word-list condition (initial rise followed by a decay). Finally, the remaining ~10% of the electrodes showed no build-up of activity and a strong degree of locking to word/nonword onsets across all conditions. Clusters 1 and 3 replicated in an independent dataset that only included two of the four linguistic conditions (sentences and nonwords); that dataset also contained a cluster that was qualitatively similar to Cluster 2 in Dataset 1.

Importantly, these findings provide evidence for ***functional heterogeneity*** within the language network. The experimental design adopted in the current study has traditionally been used as a way to tease apart neural responses to word meanings (present in sentences and lists of words, but not in Jabberwocky sentences and lists of nonwords) and syntactic structure (present in sentences and, under some views of syntax, in Jabberwocky sentences, but not in lists of words/nonwords; Fedorenko et al., 2010, 2012, 2016; for earlier uses of this paradigm, see Mazoyer et al., 1993; Friederici et al., 2000; Humphries et al., 2001; Vandenberghe et al., 2002; for another variant, see Bautista and Wilson, 2016). As measured with fMRI, all areas of the language network show a profile that suggests sensitivity to both word meanings and syntactic structure: the response is strongest to sentences, lower to word-lists and Jabberwocky

sentences, and lowest to nonword-lists (e.g., Fedorenko et al., 2010; Bedny et al., 2011; Shain et al., 2021; see Dick et al., 2001 for earlier arguments against the lexical/syntactic dissociation). Using a similar design in an intracranial recording study, Fedorenko et al. (2016) showed that the overall pattern of response to sentences, word-lists, Jabberwocky sentences, and nonword-lists mirrors that observed in fMRI studies, with no electrodes showing selective responses to lexical or syntactic processing. However, here we show that in spite of strong integration between lexical and syntactic processing, neural populations within the language network do differ functionally, although along a different dimension.

Fedorenko et al. (2016) reported a temporal profile—present in a subset of electrodes—whereby high gamma power response increases over the course of the sentence but not in other conditions (replicated by Nelson et al., 2017), which they interpreted as indexing the process of constructing a sentence-level meaning. Here, we investigated the temporal profiles of language-responsive electrodes in a more comprehensive manner. We identified language-responsive electrodes in the same way as in Fedorenko et al. (2016; building on fMRI work in Fedorenko et al., 2010), as responding more strongly to sentences than nonword-lists. However, we used a more liberal threshold to include as many potentially relevant electrodes as possible and we leveraged the fine-grained temporal information in the signal by considering the full timecourses (cf. averaging high gamma power in each word/nonword as in Fedorenko et al., 2016). In this way, we found that the build-up electrodes reported in Fedorenko et al. (2016) likely represent a mix of electrodes. The timecourse of response to the S condition in Fedorenko et al. is most similar to that in Cluster 1 here. However, the fact that in Fedorenko et al. a reliable S>W>J>N profile was present suggests a contribution from Cluster 2 electrodes. In particular, in our study, Cluster 1 electrodes show a pattern of S>W>=J=N but Cluster 2 electrodes show the S>W>J>N pattern, as in Fedorenko et al. As such, our analyses replicate the previous finding of the build-up effect (including in a new, larger dataset: Dataset 2), but identify two functionally distinct build-up profiles and uncover a third profile that does not show build-up of activity over time.

Different electrodes vary in the size of their temporal receptive windows

What do the different temporal response profiles reflect? A construct that has been steadily gaining popularity in human neuroscience is that of temporal receptive windows (TRWs). TRWs denote the amount of the preceding context that a given neural unit is sensitive to or integrates over (e.g., Hasson et al., 2008; Lerner et al., 2011; Norman-Haignere et al., 2022). Past fMRI studies have shown that the TRW of the language system is somewhere between a word and a sentence (e.g., Lerner et al., 2011; Jacoby and Fedorenko, 2020; Blank and Fedorenko, 2020; Jain et al., 2020; Caucheteux et al., 2023), although some recent work suggests that the language network may even be sensitive to sub-lexical regularities (Regev et al., 2021). Here, we argue that the observed response profiles correspond to neural populations within the language network that integrate information over different timescales, from sub-lexical units and single words (Cluster 3) to short phrases (Cluster 2) to longer phrases/sentences (Cluster 1).

We formalized this intuition through a simple model wherein we convolved a stimulus train with kernels of different widths (from 1/3 of a word to 8 words). This procedure produced simulated

neural signals that exhibit striking similarity to our observed timecourses for the sentence condition (**Figure 5A**). We defined an electrode's TRW as the kernel width that produced the simulated timecourse that correlated most strongly with the electrode's actual timecourse. This analysis suggests that the average TRW of the Cluster 3 electrodes is the shortest (~1 word), followed by the TRW of the Cluster 2 electrodes (~4 words), followed by the TRW of the Cluster 1 electrodes (~6 words, **Figure 5B, C**). Our modeling results thus indicate that TRW size is an important dimension of variation for the neural populations that comprise the language network.

Two other features of the response profiles support the idea that these response profiles correspond to different TRWs. The first has to do with how well different sets of electrodes discriminate among linguistic conditions based on response magnitude. In particular, an electrode that only processes information over the span of ~a single word (or less than a word) should show little sensitivity to whether nearby words can be composed into phrases, which should translate into weak ability to discriminate between the S and W conditions across the stimulus. This is the pattern we see for electrodes in Cluster 3 (**Figure 5A**): these electrodes do not reliably discriminate between the S and W conditions, suggesting that these electrodes do not participate in combining words into larger units. In contrast, an electrode that processes information over spans of multiple words should show sensitivity to the composability of nearby words, and thus should be able to discriminate between S and W. This is the pattern we see for electrodes in Clusters 1 and 2: these electrodes reliably discriminate between the S and W conditions.

Furthermore, electrodes that integrate information over longer spans should be more sensitive to whether or not a word can start a phrase or a sentence because such electrodes will have access to more of the preceding context (e.g., the end of the preceding trial, which is a clear clue that the next element should start a stimulus). Electrodes in Cluster 1 can already discriminate between the sentence and word-list conditions at the first word of the stimulus, presumably because the words that occur in the first position of the sentence trials can always (by definition) start a phrase/sentence, but words that occur in the first position of the word-list trials often cannot. For example, consider the sample word-list trials in **Figure 2B**: one starts with "RAIN", which *could* continue as a sentence ("Rain is sorely needed" or "Rain gear is critical in this weather") but another starts with "STOOD", which is unlikely to start a phrase/sentence. Electrodes in Cluster 2 cannot discriminate between the sentence and word-list conditions until the second or third word, which may suggest that this length of input is needed to accumulate evidence that the words in word-list trials cannot be combined (after which point the gamma response declines; **Figures 3** and **5**). For example, in a string like "RAIN THE" it is quite clear at "the" that the words are probably not combinable, but in a word-list string like "STOOD THE", it may take until the third word ("TIED") to figure out that the words cannot be combined because "STOOD THE" is compatible with continuations like "stood the test of time" or "stood the watch".

The second feature of the clusters that supports their differentiation in terms of the size of their TRWs is the degree to which their responses are locked to individual word/nonwords in the stimulus. An electrode that only processes information over the span of a single word should exhibit modulation at the rate of stimulus presentation, reflecting the momentary stimulus-

related fluctuations. On the other hand, an electrode that processes information over spans of multiple words should exhibit a response that reflects a more smoothed version of the stimulus train, with no momentary stimulus-related fluctuations. Indeed, the three clusters differ significantly in their degree of locking. Cluster 3 was the most strongly locked, followed by Cluster 2, with Cluster 1 showing the weakest amount of locking (**Figure 5C**).

As discussed earlier, prior studies have characterized the TRW of the language network as falling between a word and a sentence (e.g., Lerner et al., 2011; Jacoby and Fedorenko, 2020; Blank and Fedorenko, 2020; Shain et al., 2021). Here, we demonstrate the existence of several distinct TRWs in this range in language-responsive neural populations. The use of an extensively validated functional localizer (Fedorenko et al., 2010) to identify these language-responsive populations ensures that the observed differences reflect heterogeneity within the language system proper rather than between the language areas and nearby functionally distinct brain regions, like speech areas (e.g., Overath et al., 2015) or higher-level cognitive networks (e.g., Braga et al., 2020; Shain, Paunov, Chen et al., 2022).

It is important to note that the division of labor within the language system in terms of the scale of temporal integration is largely, though not fully, orthogonal to the putative dissociation between lexical and syntactic processing (discussed in the Introduction and above). In particular, electrodes that reliably discriminate between the word-list and nonword-list conditions (i.e., electrodes in Clusters 1 and 2; **Figure 5B**)—and are thus sensitive to meanings of individual words—all participate in combinatorial processing, integrating across words over shorter (Cluster 2) and longer (Cluster 1) spans. And neither electrodes in Cluster 2 nor in Cluster 1 show responses to linguistic structure that is independent of word meaning processing (as would be evidenced by similar profiles for sentences and Jabberwocky sentences). These results provide further support for strong integration between the processing of word meanings and phrase-structure building as has been argued in past fMRI work (e.g., Fedorenko et al., 2020). This integration likely reflects the fact that many/most inter-word dependencies in natural language are highly dependent on particular words (cf. on broad syntactic categories like nouns and verbs; e.g., Bybee, 1999, 2013; Goldberg, 2003; Jackendoff, 2007; Arnon and Snider, 2010; Jackendoff and Audring, 2020).

Discrete clusters versus a gradient of TRWs?

Do the observed response profiles reflect categorically distinct clusters that integrate information over different timescales or is the underlying structure of language-selective responses in the brain best described by a continuum of TRWs with no sharp boundaries or groupings of response types? Although we do not rule out the possibility of a TRW continuum, we highlight two aspects of the data that point to the discrete cluster architecture. First, the electrodes that exhibit the most reliable responses to linguistic input (as estimated by correlating responses to odd and even trials) are the most clearly separable in a low dimensional space (**Figure 3C**). That is, the electrodes that contribute to the appearance of a continuum tend to be more noisy. And second, the distribution (or density) of electrode TRWs (**Figure 4C**) is not uniform as we would expect if the electrodes indeed fall along a continuum. Instead, we observe

peaks/swells in the distribution (e.g., many individual electrodes are assigned a TRW of 8 words, very few are assigned a TRW of 5-6 words). This non-uniformity may suggest that the existence of neural units sensitive to information chunks of distinct and specific size is critical for efficient extraction of meaning from word sequences.

An issue that may bear on whether the language processing architecture is best characterized by a few neural populations each with a discrete TRW or by a continuum of TRWs has to do with the units in which we measure TRWs. We have so far discussed TRWs in terms of the number of words. However, TRWs may instead be bounded by the number of bits of information, which depends on how predictable words are in context (e.g., Shannon, 1949) and strongly affects behavioral (e.g., Levy, 2008) and neural (e.g., Shain, Blank et al., 2020) processing. Future work should consider multiple construals of the units in which TRWs are measured.

The distributed nature of language processing

There is a long history in language neuroscience to try to separate the process of language comprehension into stages that are not only temporally distinct but are also carried out in spatially distinct brain areas. At some level, this kind of architecture indeed characterizes language processing. In particular, clear separation exists between the language-processing system (Fedorenko et al., 2011) and both a) lower-level perceptual areas, like the speech-perception areas (Norman-Haignere et al., 2015; Overath et al., 2015) or the visual word-form area (e.g., Baker et al., 2007; Hamamé et al., 2013; Saygin et al., 2016) and b) putative higher-level areas (like the Default network; Buckner & DiNicola, 2019; or the Theory of Mind network; Saxe et al., 2006) that may carry out further processing on the meaning representations extracted from language (e.g., Baldassano et al., 2017; 2018). However, finding spatial sub-divisions within the language-selective network has proven challenging (e.g., Fedorenko et al., 2010, 2020; Bautista & Wilson, 2016; Regev et al., 2021).

The current work demonstrates that although there exist functional differences within the language network, these functionally distinct mechanisms are *not spatially clustered* and are instead distributed in an interleaved fashion across the language network. The latter explains why most past fMRI work could not discover this functional heterogeneity (cf. Fedorenko et al., 2012 for implied functional heterogeneity given the information present in multivariate patterns of fMRI response; and see Jain et al., 2020 for evidence of voxel-level heterogeneity with respect to TRWs as discovered in an encoding approach with artificial neural network (ANN) language models).

Future directions

The current findings lay the foundation for several exciting future research avenues. First, the size of a neural unit's temporal receptive window (TRW) should determine its sensitivity to different linguistic features. For instance, electrodes with short TRWs, but not with longer ones should be sensitive to features like word length and lexical frequency; in contrast, only electrodes with long TRWs should be sensitive to the difficulty of forming non-local dependencies.

Furthermore, our TRW modeling was performed only on the neural responses to the sentence condition and not the linguistically degraded conditions. Future work should incorporate detailed linguistic analysis of the input to arrive at a more complete neural model of language comprehension.

Second, artificial neural network (ANN) language models could be leveraged to better understand the constraints on language processing architecture. For example, do successful language architectures require particular proportions of units with different TRWs or particular distributions of such units within and/or across model layers? In Dataset 1, we found fewest electrodes belonging to Cluster 3 (shortest TRW), more electrodes belonging to Cluster 2 (intermediate TRW), and the majority of electrodes belonging to Cluster 1 (longest TRW). These proportions align with the idea that compositional semantic space is highly multi-dimensional, but word-form information can be represented in a relatively low-dimensional space (e.g., Mollica and Piantadosi, 2019). However, the proportions can also be affected by biases in where intracranial electrodes tend to be implanted, so investigating these questions in ANNs, where we can probe all units in the network and have the freedom to alter the architecture in various ways, may yield insights that cannot be gained from human brains at least with the current experimental tools available.

And third, much recent evidence suggests that human comprehension mechanisms are robust to noise in the input (e.g., Levy, 2008; Gibson et al., 2013; Gibson et al., 2017; Keshev & Meltzer-Asscher, 2021; Ryskin et al., 2018, 2021; see Gibson et al., 2019 for a review). This property of the language processing system may make it desirable for TRWs to be somewhat flexible, so as to allow for the possibility of revising interpretation with incoming input. Understanding how this feat is accomplished—for example, whether any given neural population's TRW is not strictly fixed or whether this flexibility is accomplished by multiple neural populations working together—will require a deeper understanding of the dynamics of information processing in the neural populations with different TRWs and careful manipulations of stimulus properties.

A key limitation of this work is that we have focused on the similarity of temporal dynamics across sentences (and other linguistic stimuli) regardless of their structure and meaning. However, in order for the language system to extract meaning from the signal, language-responsive electrodes have to represent specific linguistic content. Advances in the recording technologies available for human neuroscience (e.g., Paulk et al., 2022), combined with the Increasing use of ANN language models for understanding the human language system (Toneva and Wehbe, 2019; Jain et al., 2020; Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux & King, 2022) may allow us to soon go beyond this coarse-level functional differentiation and to start uncovering the process of building particular linguistic meanings.

Methods

Participants

Dataset 1 (also used in Fedorenko et al., 2016)*:* Electrophysiological data were recorded from intracranial electrodes in 6 participants (5 female, aged 18–29 years) with intractable epilepsy. These participants underwent temporary implantation of subdural electrode arrays at Albany Medical Center to localize the epileptogenic zones and to delineate it from eloquent cortical areas before brain resection. All participants gave informed written consent to participate in the study, which was approved by the Institutional Review Board of Albany Medical Center. One further participant was tested but excluded from analyses because of difficulties in performing the task (i.e., pressing multiple keys, looking away from the screen) during the first five runs. After the first five runs, the participant required a long break during which a seizure occurred.

Dataset 2: Electrophysiological data were recorded from intracranial electrodes in 16 participants (4 female, aged 21-66 years) with intractable epilepsy. These participants underwent temporary implantation of subdural electrode arrays and depth electrodes to localize the epileptogenic zones before brain resection at one of four sites: Albany Medical Center (AMC), Barnes-Jewish Hospital (BJH), Mayo Clinic Jacksonville (MCJ), and St. Louis Children's Hospital (SLCH). All participants gave informed written consent to participate in the study, which was approved by the Institutional Review Board at each relevant site. Two further participants were tested but excluded from analyses due to the lack of any language-responsive electrodes (see Language-Responsive Electrode Selection).

Data Collection

Dataset 1: The implanted electrode grids consisted of platinum-iridium electrodes that were 4 mm in diameter (2.3–3 mm exposed) and spaced with an inter-electrode distance of 0.6 or 1 cm. The total numbers of implanted grid/strip electrodes were 120, 128, 98, 134, 98, and 36 for the six participants, respectively (**Table 1**). Electrodes were implanted in the left hemisphere for all participants except P6, who had bilateral coverage (16 left hemisphere electrodes). Signals were digitized at 1,200 Hz.

Dataset 2: The implanted electrode grids and depth electrodes consisted of platinum-iridium electrodes. Implanted grid contacts were spaced at 0.6 or 1cm (2.3–3 mm exposed), while SEEG leads were spaced 3.5 - 5 mm depending on the trajectory length, with 2 mm exposed. The total numbers of implanted electrodes by participant can be found in **Table 2** (average=167 electrodes; st. dev.=51; range 92-234), along with the frequencies at which the signals were digitized. Electrodes were implanted in only the left hemisphere for 2 participants, in only the right hemisphere for 2 participants, and bilaterally for 12 participants (**Table 2**). All participants, regardless of the lateralization of their coverage, were included in all analyses.

For both datasets, recordings were synchronized with stimulus presentation and stored using the BCI2000 software platform (Schalk et al., 2004).

Cortical Mapping

Electrode locations were obtained from post-implantation computerized tomography (CT) imaging and co-registered with the 3D surface model of each participant's cortex—created from the preoperative anatomical MRI image—using the VERA software suite (Adamek et al., 2022). Electrode locations were then transformed to MNI space within VERA via nonlinear co-registration of the subjects' skull-stripped anatomical scan and the skull-stripped MNI152 Freesurfer template using ANTs (Avants, Epstein, Grossman, & Gee, 2008).

Preprocessing and Extraction of Signal Envelope

Neural recordings were collected and saved in separate data files by run (see Experiment and **Tables 1-2**), and all preprocessing procedures were applied *within* data files to avoid inducing artifacts around recording breaks.

First, the ECoG/sEEG recordings were high-pass filtered at the frequency of 0.5 Hz, and line noise was removed using IIR notch filters at 60, 120, 180, and 240 Hz. The following electrodes (electrodes; we use these terms interchangeably) were excluded from analysis: a) ground, b) reference, and c) those that were not ECoG or sEEG contacts (e.g., microphone electrodes, trigger electrodes, scalp electroencephalography (EEG) electrodes, EKG electrodes), as well as d) those with significant line noise, defined as electrodes with line noise greater than 5 standard deviations above other electrodes, e) those with large artifacts identified through visual inspection, and, for all but four participants, f) those that had a significant number of interictal discharges identified using an automated procedure (Janca et al., 2015). (For 4 participants—P3 in Dataset 1 and P15, P17, and P21 in Dataset 2—electrodes that were identified as having a significant number of interictal discharges were not excluded from analyses because more than 1/3 of each of these participants' electrodes fit this criterion.) These exclusion criteria left 108, 115, 92, 106, 93, and 36 electrodes for analysis for the 6 participants in Dataset 1 (**Table 1**) and between 76 and 228 electrodes for the 16 participants in Dataset 2 (**Table 2**).

Next, the common average reference (from all electrodes connected to the same amplifier) was removed for each timepoint separately. The signal in the high gamma frequency band (70 Hz–150 Hz) was then extracted by taking the absolute value of the Hilbert transform of the signal extracted from 8 gaussian filters (center frequencies: 73, 79.5, 87.8, 96.9, 107, 118.1, 130.4, and 144; standard deviations (std): 4.68, 4.92, 5.17, 5.43, 5.7, 5.99, 6.3, and 6.62, respectively, as in e.g., Dichter et al., 2018). The resulting envelopes from each of the Gaussian filters were averaged into one high gamma envelope. We focus on the high gamma frequency range because this component of the signal has been shown to track neural activity most closely (e.g., Janca et al., 2015). Linear interpolation was used to remove data points whose magnitude was more than 5 times the 90^{th} percentile of all magnitudes (Norman-Haignere et al., 2022), and we downsampled the signal by a factor of 4. For all data analysis basic Matlab (version 2021a) functions were used.

Finally, the data were z-scored and normalized to a min/max value of 0/1 to allow for comparisons across electrodes, and the signal was downsampled further to 60 Hz (regardless of the participant's native sampling frequency) to reduce noise and standardize the sampling frequency across participants. For the participants who performed a slower version of the

paradigm (e.g., words presented for 700 ms each; see Experiment), the signal was time-warped to a faster rate (words presented for 450 ms each) so that timecourses could be compared across subjects.

Experiment

Dataset 1: In an event-related design, participants read sentences, lists of words, Jabberwocky sentences, and lists of nonwords. The materials were adapted from Fedorenko et al. (2010; Experiment 2). Each event (trial) consisted of eight words/nonwords, presented one at a time at the center of the screen. At the end of each sequence, a memory probe was presented (a word in the sentence and word-list conditions, and a nonword in the Jabberwocky and nonword-list conditions) and participants had to decide whether the probe appeared in the preceding sequence by pressing one of two buttons. Two different presentation rates were used: P1, P5, and P6 viewed each word/nonword for 450 ms (fast-timing), and P2, P3, and P4 viewed each word/nonword for 700 ms (slow-timing). The presentation speed was determined before the experiment based on the participant's preference. After the last word/nonword in the sequence, a fixation cross was presented for 250 ms, followed by the probe item (1,400-ms fast-timing, 1,900 ms slow-timing), and a post-probe fixation (250 ms). Behavioral responses were continually recorded. After each trial, a fixation cross was presented for a variable amount of time, semi-randomly selected from a range of durations from 0 to 11,000 ms, to obtain a low-level baseline for neural activity.

Trials were grouped into runs to give participants short breaks throughout the experiment. In the fast-timing version of the experiment, each run included eight trials per condition and lasted 220 s, and in the slow-timing version, each run included six trials per condition and lasted 264 s. The total amount of intertrial fixation in each run was 44 s for the fast-timing version and 72 s for the slow-timing version. All participants completed 10 runs of the experiment, for a total of 80 trials per condition in the fast-timing version and 60 trials per condition in the slow-timing version.

Dataset 2: In an event-related design that was similar to the one used in Dataset 1, participants read sentences and lists of nonwords (the other two conditions—lists of words and Jabberwocky sentences—were not included). The materials were adapted from a version of the language localizer in use in the Fedorenko lab (e.g., Malik-Moraleda, Ayyash et al., 2022). Each event (trial) consisted of twelve words/nonwords, presented one at a time at the center of the screen. At the end of each sequence, a memory probe was presented (a word in the sentence condition and a nonword in the nonword-list condition) and participants had to decide whether the probe appeared in the preceding sequence by pressing one of two buttons. Two presentation rates were used: 600 ms per word/nonword (medium-timing) and 750 ms per word/nonword (slow-timing; see **Table 2** for a description of the presentation rates by participant). The presentation speed was determined before the experiment based on the participant's preference. After the last word/nonword in the sequence, a fixation cross was presented for 400 ms, followed by the probe item (1,000 ms for both fast- and slow-timing), and a post-probe fixation (600 ms). Behavioral responses were continually recorded. After each trial, a fixation cross was presented for a variable amount of time, semi-randomly selected from a range of durations from 0 to 6,000 ms.

Trials were grouped into runs to give participants short breaks throughout the experiment. In the medium-timing version of the experiment, each run included 36 trials per condition and lasted ~898 s, and in the slow-timing version, each run included 24 trials per condition and lasted 692 s. The total amount of intertrial fixation in each run was 216 s for the medium-timing version and 144 s for the slowest-timing version. One participant (P7) saw a modified slow-timing version of the paradigm where only 48 of the full 72 items per condition were shown. 13 participants completed 2 runs of the experiment (all saw the medium-timing version, 72 trials per condition), 2 participants completed 3 runs of the experiment (one saw the slow-timing version, 72 trials per condition; and the other saw the modified slow-timing version, 48 trials per condition), and 1 participant completed 1 run of the experiment (medium-timing version, 36 trials per condition, **Table 2**).

For all clustering analyses, only the first eight words/nonwords of the stimulus were used to ensure that the length of the timecourses being analyzed was the same across Dataset 1 and 2.

Language-Responsive Electrode Selection

In both datasets, we identified language-responsive electrodes as electrodes that respond significantly more (on average, across trials) to sentences (the S condition) than to perceptually similar but linguistically uninformative (i.e., meaningless and unstructured) nonword sequences (the N condition). First, the envelope of the high-gamma signal was averaged across word/nonword positions (8 positions in the experiment used in Dataset 1, and 12 positions in the experiment used in Dataset 2) to construct an 'observed' response vector for each electrode (1 x nTrialsS + nTrialsN; the number of trials, across the S and N conditions, varied by participant between 72 and 160). The observed response vector was then correlated (using Spearman's correlation) with an 'idealized' language response vector, where sentence trials were assigned a value of 1 and nonword trials—a value of -1. The values in the ideal response vector were then randomly permuted without replacement and a new correlation was computed. This process was repeated 10,000 times, for each electrode separately, to construct a null distribution (with shuffled labels) relative to which the true correlation between the observed values and the 'idealized' values could be evaluated. Electrodes were determined to be language-responsive if the observed vs. idealized correlation was greater than 95% of the correlations computed using the permuted idealized response vectors (equivalent to $p < 0.05$). (We chose a liberal significance threshold in order to maximize the number of electrodes to be included in the critical analyses, and to increase the chances of discovering distinct response profiles.) The majority of the language-responsive electrodes (98.3% in Dataset 1, 53.9% in Dataset 2) fell in the left hemisphere, but we use electrodes across both hemispheres in all analyses (see e.g., Lipkin et al., 2022 for evidence of a robust right-hemisphere component of the language network in a dataset of >800 participants).

Participants	Age	Sex	Site	ECoG or sEEG	Language-responsive electrodes (S>N)	Total clean electrodes	Total electrodes	Native sampling freq (Hz)	Elec per amp	Runs	Pres. rate (per word)	Trials per cond
Participant 1	29	F	AMC	ECoG	62 (0 RH)	108 (0 RH)	120 (0 RH)	1200	16	10	450ms	80
Participant 2	25	F	AMC	ECoG	17 (0 RH)	115 (0 RH)	128 (0 RH)	1200	16	10	700ms	60
Participant 3	18	F	AMC	ECoG	17 (0 RH)	92 (0 RH)	98 (0 RH)	1200	16	10	700ms	60
Participant 4	28	M	AMC	ECoG	26 (0 RH)	106 (0 RH)	134 (0 RH)	1200	64	10	700ms	60
Participant 5	25	F	AMC	ECoG	48 (0 RH)	93 (0 RH)	98 (0 RH)	1200	64	10	450ms	80
Participant 6	20	F	AMC	ECoG	7 (3 RH)	36 (20 RH)	36 (20 RH)	1200	64	10	450ms	80

Table 1: Details for Dataset 1. (All data were collected at the Albany Medical Center (Site=AMC).) Here and in Table 2, 'Total electrodes' excludes reference electrodes, ground electrodes, microphone electrodes, trigger electrodes, skull EEG electrodes, and EKG electrodes; and 'Total clean electrodes' excludes electrodes with significant line noise, significant interictal discharges, or large visual artifacts identified through manual inspection. 'Elec per amp' – Number of electrodes per amplifier. 'Pres rate (per word)' – duration of presentation of each single word or nonword.

Participant	Age	Sex	Site	ECoG or sEEG	Language-responsive electrodes (S>N)	Total clean electrodes	Total electrodes	Native sampling freq (Hz)	Elec per amp	Runs	Pres rate (per word)	Trials per cond
Participant 7	51	M	AMC	EcoG	14 (7 RH)	116 (25 RH)	126 (26 RH)	1200	64	3	750ms	48
Participant 8	30	F	AMC	both	18 (0 RH)	76 (1 RH)	92 (3 RH)	1200	64	3	750ms	72
Participant 9	31	M	AMC	sEEG	2 (1 RH)	90 (44 RH)	98 (52 RH)	1200	64	2	600ms	72
Participant 10	59	F	AMC	sEEG	2 (0 RH)	113 (0 RH)	124 (0 RH)	1200	64	2	600ms	72
Participant 11	23	M	AMC	EcoG	58 (33 RH)	209 (110 RH)	216 (110 RH)	1200	64	2	600ms	72
Participant 12	39	M	AMC	sEEG	5 (5 RH)	112 (112 RH)	128 (128 RH)	1200	64	2	600ms	72
Participant 13	29	M	AMC	EcoG	9 (0 RH)	126 (0 RH)	132 (0 RH)	1200	64	2	600ms	72
Participant 14	36	M	AMC	sEEG	3 (2 RH)	169 (84 RH)	184 (90 RH)	1200	64	2	600ms	72
Participant 15	25	M	BJH	sEEG	19 (16 RH)	183 (93 RH)	183 (93 RH)	1000	64	2	600ms	72
Participant 16	38	M	BJH	sEEG	49 (15 RH)	169 (72 RH)	224 (112 RH)	1000	64	2	600ms	72
Participant 17	31	F	BJH	sEEG	17 (0 RH)	228 (30 RH)	228 (30 RH)	1000	64	2	600ms	72
Participant 18	40	M	BJH	sEEG	35 (5 RH)	137 (11 RH)	192 (14 RH)	1000	64	2	600ms	72
Participant 19	66	M	BJH	sEEG	32 (1 RH)	210 (13 RH)	234 (16 RH)	2000	64	2	600ms	72
Participant 20	24	M	BJH	sEEG	7 (0 RH)	156 (30 RH)	218 (30 RH)	2000	64	2	600ms	72
Participant 21	39	M	MCJ	sEEG	11 (1 RH)	108 (45 RH)	109 (45 RH)	1200	64	1	600ms	36
Participant 22	21	F	SLCH	sEEG	81 (81 RH)	176 (176 RH)	186 (186 RH)	2000	64	2	600ms	72

Table 2: Details for Dataset 2. (The data were collected at four sites: Albany Medical Center (Site=AMC), Barnes-Jewish Hospital (Site=BJH), Mayo Clinic Jacksonville (Site=MCJ), and St. Louis Children's Hospital (Site=SLCH)).

Clustering analysis

Using Dataset 1 (n=6 participants, m=177 language-responsive electrodes), we created a single timecourse per electrode by concatenating the average timecourses across the four conditions (sentences (S), word-lists (W), Jabberwocky sentences (J), nonword-lists (N)). All the timepoints of the concatenated timecourses (864 data points: 60 Hz * 4 conditions * 3.60 seconds per trial after resampling) served as input to a k-medoids clustering algorithm (Kaufman & Rousseuw, 1990). K-medoids is a clustering technique that divides data points—electrodes in our case—into k groups, where k is predetermined. The algorithm attempts to minimize the distances between each electrode and the cluster center, where cluster centers are represented by 'medoids' (exemplar electrodes selected by the algorithm) and the distance metric is correlation-based. K-medoids clustering was chosen over the more commonly used k-means clustering to allow for

the use of a correlation-based distance metric as we were most interested in the shape of the timecourses over their scale which can vary due to cognitively irrelevant physiological differences (but see **Figure S1** for evidence that similar clusters emerge with a k-means clustering algorithm using a Euclidean distance).

Optimal number of clusters

To determine the optimal number of clusters, we used the "elbow" method (e.g., Rokach and Maimon, 2005) which searches for the value of k above which the increase in explained variance becomes more moderate. For each k (between 2 and 10), k-medoids clustering was performed, and explained variance was computed as the sum of the correlation-based distances of all the electrodes to their assigned cluster center and normalized to the sum of the distances for k=1 (equivalent to the variance of the full dataset). This explained variance was plotted against k and the "elbow" was determined as the point after which the derivative became more moderate. We also plotted the derivative of this curve for easier inspection of the transition point.

Electrode discrimination between conditions

To examined the *timecourse of condition divergence*, as quantified by the electrodes' ability to linearly discriminate between the magnitudes of pairs of conditions. We focused on condition pairs that critically differ in their engagement of particular linguistic processes: conditions S and W, which differ in whether they engage combinatorial (syntactic and semantic) processing (S=yes, W=no), and conditions W and N, which differ in whether they engage word meaning processing (W=yes, N=no). This analysis tests how early the relevant pair of conditions reliably diverge. For every electrode, the mean response to the three conditions of interest (S, W, and N) was averaged across 100 ms bins with a 50 ms sliding window (i.e., 50% overlap between adjacent time bins). For each cluster separately, a set of 20 models (binary logistic classifiers) was trained (to discriminate S from W, or W from N) at each time bin using the binned neural signal up to, and including, that time bin. Each classifier was trained using 10-fold cross validation (train on 90% of the data and test using the remaining 10%, repeated for 10 splits of the data such that every observation was included in the test set exactly once). The predicted and actual conditions across all folds were used to calculate one accuracy per classifier (the percent of mean responses from all electrodes in a particular cluster correctly classified as S/W or W/N; 20 accuracies in total). The performance of the set of models at a given time bin was evaluated against the 50% (chance) baseline using a one-sample t-test.

Electrode locking to onsets of individual word/nonwords

To estimate the degree of stimulus locking for each electrode and condition separately, we fitted a sinusoidal function that represented the stimulus train to the mean of the odd trials and then computed the Pearson correlation between the fitted sinusoidal function and the mean of the even trials. For the sinusoidal function fitting, we assumed that the frequency of the sinusoidal function was the frequency of stimulus presentation and we fitted the phase, amplitude and offset of the sinusoidal by searching parameter combinations that minimized the sum of squared differences between the estimated sinusoidal function and the data. Cross-validation (fitting on odd trials and computing the correlation on even trials) ensured non-circularity. To statistically quantify differences in the degree of stimulus locking between the clusters and among the

conditions, we ran a linear mixed-effects (LME, using the MATLAB procedure *fitlme*: MATLAB) model regressing the locking values of all electrodes and conditions on the fixed effects categorical variable of *cluster* (with 3 levels for Cluster 1, 2 or 3 according to which cluster each electrode was assigned to) and *condition* (with 4 levels for conditions S, W, J, N), both grouped by the random effects variable of *participant*, as well as a fixed interaction term between *cluster* and *condition*:

*Locking ~ 1 + cluster*condition + (cluster|participant) + (condition|participant)*

An ANOVA test for LME was used to determine the main effects of *cluster* and *condition* and their interaction. Pairwise comparisons of all 3 clusters and 4 conditions as well as interactions between all (cluster, condition) pairs were extracted from the model estimates.

Cluster stability across trials

We evaluated the stability of the clustering solution by performing the same clustering procedure as described above (Clustering analysis) on half the trials. To evaluate the similarity of the clusters derived based on only half of the trials to the clusters derived based on all trials, we first had to determine how clusters correspond between any two solutions. In particular, given that the specific order of the clusters that the k-medoids algorithm produces depends on the (stochastic) choice of initial cluster medoids, the electrodes that make up cluster 1 in one solution may be labelled as cluster 2 in another solution. To determine cluster correspondence across solutions, we matched the cluster centers (computed here as the average timecourse of all electrodes in a given cluster) from a solution based on half of the trials to the most highly correlated cluster centers from the solution based on all trials.

We then conducted a permutation analysis to statistically compare the clustering solutions. This was done separately for each of the two halves of the data (odd- and even-numbered subsets of trials). Under the null hypothesis, no distinct response profiles should be detectable in the data, and consequently, responses in one electrode should be interchangeable with responses in another electrode. Using half of the data, we shuffled individual trials across electrodes (within each condition separately), re-clustered the electrodes into 3 clusters, and then correlated the resulting cluster centers to the 'corresponding' cluster centers from the full dataset. We repeated this process 1,000 times to construct a null distribution of the correlations for each of the 3 clusters. These distributions were used to calculate the probability that the correlation between the clusters across the two solutions using the actual, non-permuted data was higher than would be expected by chance.

This permutation analysis was used in all subsequent analyses that compare two clustering solutions (e.g., comparing clusters from Dataset 2 with clusters from Dataset 1).

Cluster robustness to data loss

We evaluated the robustness of the clustering solution to loss of electrodes to ensure that the solution was not driven by particular electrodes or participants.

To evaluate the similarity of the clusters derived based on only a subset of language-responsive electrodes to the clusters derived based on all electrodes, we progressively removed electrodes from the full set (n=177) until only 3 electrodes remained (required to split the data into 3 clusters) in increments of 5. Each subset of electrodes was clustered into 3 clusters, and the cluster centers were correlated with the corresponding cluster centers (see 1 above) from the full set of electrodes. For each subset of electrodes, we repeated this process 100 times, omitting a different random set of n electrodes with replacement, and computed the average correlation across repetitions.

To statistically evaluate whether the clustering solutions with only a subset of electrodes were more similar to the solution on the full set of electrodes on average (across the 100 repetitions at each subset size) than would be expected by chance, we conducted a permutation analysis like the one described in 1. In particular, using the full dataset, we shuffled individual trials across electrodes (within each condition separately), re-clustered the electrodes into 3 clusters, and then correlated the resulting cluster centers to the 'corresponding' cluster centers from the actual, non-shuffled data. We repeated this process 1,000 times to construct a null distribution of the correlations for each of the 3 clusters. These distributions were used to calculate the probability that the correlation between the clusters across the two solutions using the actual, non-permuted data (a solution on all the electrodes and a solution on a subset of the electrodes) was higher than would be expected by chance. To err on the conservative side, we chose the null distribution for the cluster with the highest average correlation in the permuted version of the data. For each subset of electrodes, if the average correlation (across the 100 repetitions) fell below the 95th percentile of the null distribution, this was taken to suggest that the clustering solution based on a subset of the electrodes was no longer more correlated to the solution on the full set of electrodes than would be expected by chance.

Estimation of temporal receptive window size per electrode

We used a simplified model to simulate neural responses in the sentence (S) condition as a convolution of a stimulus train and gaussian kernels with varying widths. The kernels represented the temporal receptive window (TRW) of an idealized neural population underlying the intracranial responses measured by a single electrode. The kernels were constructed from gaussian curves with a standard deviation of $\sigma/2$ truncated at +/- 1 standard deviation (capturing 2/3 of the area under the gaussian, **Figure 5A**). We then normalized the truncated gaussians to have a minimum of 0 and maximum of 1. The stimulus train was represented by 1 at the time of new word onsets and 0 otherwise. The resulting simulated neural signals were also normalized to be between 0 and 1. Neural responses were simulated for σ ranging from one third of a word to 8 words (the length of our stimuli), in 1 sample increments (60Hz, 1/27th of a word, the highest resolution we were able to evaluate at the given sampling rate). To find the best fit of the receptive window size for each electrode, we selected the TRW size that yielded the highest correlation between the simulated and actual neural response. The value of the correlation was taken as a proxy for the goodness of fit.

To evaluate significance we ran linear mixed-effects (LME) models regressing the estimates temporal receptive window sizes (σ) of all electrodes on the fixed effects categorical variable of

cluster grouped by the random effects variable of *participant*. Cluster was dummy-coded as a categorical variable with three levels, and Cluster 1 was treated as the baseline intercept. This approach allowed us to compare electrodes in Cluster 2 to those in Cluster 1, and electrodes in Cluster 3 to those in Cluster 1. To additionally compare electrodes in Clusters 2 vs. 3, we compared their LME coefficients using the MATLAB procedure *coefTest*.

Anatomical topography analysis

We examined the anatomical topographic distribution of the electrodes that exhibit the three temporal response profiles discovered in Dataset 1. Specifically, we probed the spatial relationships between all pairs of electrodes that belong to different clusters (e.g., electrodes in Cluster 1 vs. 2) with respect to the two axes: anterior-posterior [y], and inferior-superior [z]. This approach allowed us to ask whether, for example, electrodes that belong to one cluster tend to consistently fall posterior to the electrodes that belong to another cluster.

To do this, we extracted the MNI coordinates of all the electrodes in each of the three clusters and ran linear mixed-effects (LME) models regressing each of the coordinates (x, y, and z) on the fixed effects categorical variable of *cluster* grouped by the random effects variable of *participant*. The random effect that groups the estimates by participant ensures that electrode coordinates are compared *within participants*, which is important given the inter-individual variability in the precise locations of language areas (e.g., Fedorenko et al., 2010), which means that the absolute values of MNI coordinates cannot be easily compared between participants. Cluster was dummy-coded as a categorical variable with three levels, and Cluster 1 was treated as the baseline intercept. This approach allowed us to compare electrodes in Cluster 2 to those in Cluster 1, and electrodes in Cluster 3 to those in Cluster 1. To additionally compare electrodes in Clusters 2 vs. 3, we compared their LME coefficients using the MATLAB procedure *coefTest*.

We repeated this analysis for Dataset 2, but we only examined Clusters 1 and 3, which were robustly present in that dataset. We performed the analysis for the electrodes in the two hemispheres separately.

Replication of the clusters in Dataset 2.

As described in Experiment, the design that was used for participants in Dataset 1 included four conditions: sentences (S), word-lists (W), Jabberwocky sentences (J), and nonword-lists (N). Because the design in Dataset 2 included only two of the four conditions (sentences (S) and nonword-lists (N)), we first repeated the clustering procedure for Dataset 1 using only the S and N conditions to test whether similar clusters could be recovered with only a subset of conditions.

We then applied the same clustering procedure to Dataset 2 (n=16 participants, m=362 language-responsive electrodes). The elbow method revealed that the optimal number of clusters in Dataset 2 is k=2. However, because the optimal number of clusters in Dataset 1 was k=3, we examined the clustering solutions at both k=2 and k=3 levels.

To statistically compare the clustering solutions between Datasets 1 and 2 for k=3, we used the same procedure as the one described above (Stability of clusters across trials). In particular, using

Dataset 2, we shuffled individual trials across electrodes (within each condition separately), re-clustered the electrodes into 3 clusters, and then correlated the resulting cluster centers to the 'corresponding' cluster centers from Dataset 1. We repeated this process 1,000 times to construct a null distribution of the correlations for each of the 3 clusters. These distributions were used to calculate the probability that the correlation between the clusters across the two datasets using the actual, non-permuted Dataset 2 was higher than would be expected by chance.

To statistically compare the clustering solutions when k=3 in Dataset 1 and k=2 in Dataset 2, we used a similar procedure as the one described above. However, we only compared the resulting cluster centers from the permuted data to the two clusters in Dataset 1 that were most strongly correlated with the two clusters that emerge from Dataset 2 (i.e., Clusters 1 and 3).

Data Availability

Preprocessed data will be publicly available on OpenNeuro at the time of publication. All stimuli will be available on OSF as well. Raw data will be made available upon request.

Code Availability

Code used to conduct analyses and generate figures from the preprocessed data will be publicly available on GitHub at the time of publication.

References

Adamek M, Swift JR, Brunner P (2022). VERA - Versatile Electrode Localization Framework.

Arnon I, Snider N (2010). More than words: Frequency effects for multi-word phrases. J Mem Lang 62:67–82.

Baker CI, Liu J, Wald LL, Kwong KK, Benner T, Kanwisher N (2007). Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. Proc Natl Acad Sci USA 104:9087–9092.

Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. Neuron 95(3):709–721.

Baldassano C, Hasson U, Norman KA (2018). Representation of Real-World Event Schemas during Narrative Perception. J Neurosci 38(45):9689–9699.

Bautista A, Wilson SM (2016). Neural responses to grammatically and lexically degraded speech. Lang Cogn Neurosci 31:567–574.

Bedny M, Pascual-Leone A, Dodell-Feder D, Fedorenko E, Saxe R (2011). Language processing in the occipital cortex of congenitally blind adults. Proc Natl Acad Sci USA 108:4429–4434.

Blank I, Balewski Z, Mahowald K, Fedorenko E (2016) Syntactic processing is distributed across the language system. Neuroimage 127:307–323.

Blank I, Kanwisher N, Fedorenko E (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. J Neurophysiol 112:1105–1118.

Blank IA, Fedorenko E (2020). No evidence for differences among language regions in their temporal receptive windows. Neuroimage 219.

Braga RM, DiNicola LM, Becker HC, Buckner RL (2020). Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. J Neurophysiol 124:1415–1448.

Buckner RL, DiNicola LM (2019). The brain's default network: updated anatomy, physiology and evolving insights. Nat Rev Neurosci 20(10):593–608.

Bybee J (1999). Usage-based Phonology. In: Functionalism and Formalism in Linguistics: Volume I: General papers, pp 211–242. John Benjamins Publishing.

Bybee J (2013). Usage-based theory and exemplar representations of constructions. In: The Oxford Handbook of Construction Grammar, pp 49–69.

Caucheteux, C., Gramfort, A. & King, JR (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. Nature Human Behaviour.

Chen X, Affourtit J, Ryskin R, Regev TI, Norman-Haignere S, Jouravlev O, Malik-Moraleda S, Kean H, Varley R, Fedorenko E (2021). The human language system does not support music processing. bioRxiv:2021.06.01.446439.

Cheung C, Ivanova A, Siegelman M, Pongos A, Kean H, Fedorenko E (2020). The effect of task on sentence processing in the brain. In: Poster Presentation at the Society for the Neurobiology of Language.

Dapretto M, Bookheimer SY (1999). Form and Content: Dissociating Syntax and Semantics in Sentence Comprehension. Neuron 24:427–432.

Deen B, Koldewyn K, Kanwisher N, Saxe R (2015). Functional organization of social perception and cognition in the superior temporal sulcus. Cereb Cortex 25:4596–4609.

DeWitt I, Rauschecker JP (2012). Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci USA 109:2709.

Diachek E, Blank I, Siegelman M, Affourtit J, Fedorenko E (2020). The domain-general multiple demand (MD) network does not support core aspects of language comprehension: A large-scale fMRI investigation. J Neurosci 40:4536–4550.

Dichter BK, Breshears JD, Leonard MK, Chang EF (2018). The Control of Vocal Pitch in Human Laryngeal Motor Cortex. Cell 174:21-31.

Dick F, Bates E, Utman JA, Wulfeck B, Dronkers N, Gernsbacher MA (2001). Language deficits, localization, and grammar: Evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. Psychol Rev 108:759–788.

Embick D, Marantz A, Miyashita Y, O'Neil W, Sakai KL (2000). A syntactic specialization for Broca's area. Proc Natl Acad Sci USA 97:6150–6154.

Fedorenko E, Behr MK, Kanwisher N (2011). Functional specificity for high-level linguistic processing in the human brain. Proc Natl Acad Sci USA 108:16428–16433.

Fedorenko E, Blank IA, Siegelman M, Mineroff Z (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. Cognition 203:104348.

Fedorenko E, Hsieh PJ, Nieto-Castañón A, Whitfield-Gabrieli S, Kanwisher N (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. J Neurophysiol 104:1177–1194.

Fedorenko E, Nieto-Castañon A, Kanwisher N (2012). Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. Neuropsychologia 50:499–513.

Fedorenko E, Scott TL, Brunner P, Coon WG, Pritchett B, Schalk G, Kanwisher N (2016). Neural correlate of the construction of sentence meaning. Proc Natl Acad Sci USA 113:E6256–E6262.

Friederici AD (2002). Towards a neural basis of auditory sentence processing. Trends Cogn Sci 6:78–84.

Friederici AD (2011). The brain basis of language processing: From structure to function. Physiol Rev 91:1357–1392.

Friederici AD, Meyer M, Von Cramon DY (2000). Auditory Language Comprehension: An Event-Related fMRI Study on the Processing of Syntactic and Lexical Information. Brain Lang 74:289–300.

Gibson E, Bergen L, Piantadosi ST (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. Proc Natl Acad Sci USA 110(20):8051-8056.

Gibson E, Tan C, Futrell R, Mahowald K, Konieczny L, Hemforth B, Fedorenko E (2017). Don't Underestimate the Benefits of Being Misunderstood. Psychol Sci 28(6):703–712.

Gibson E, Futrell R, Piantadosi SP, Dautriche I, Mahowald K, Bergen L, Levy R (2019). How Efficiency Shapes Human Language. Trends Cogn Sci 23(5):389–407.

Goldberg AE (2003). Constructions: A new theoretical approach to language. Trends Cogn Sci 7:219–224.

Goldstein A et al. (2022). Shared computational principles for language processing in humans and deep language models. Nat Neurosci 25:369–380.

Graves WW, Grabowski TJ, Mehta S, Gupta P (2008). The left posterior superior temporal gyrus participates specifically in accessing lexical phonology. J Cogn Neurosci 20:1698–1710.

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

Hagoort P (2005) On Broca, brain, and binding: a new framework. Trends Cogn Sci 9:416–423.

Hamamé CM, Szwed M, Sharman M, Vidal JR, Perrone-Bertolotti M, Kahane P, Bertrand O, Lachaux JP (2013). Dejerine's reading area revisited with intracranial EEG. Neurology 80:602–603.

Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N (2008). A Hierarchy of Temporal Receptive Windows in Human Cortex. J Neurosci 28:2539–2550.

Humphries C, Willard K, Buchsbaum B, Hickok G (2001). Role of anterior temporal cortex in auditory sentence comprehension: an fMRI study. Neuroreport 12:1749–1752.

Huth AG, De Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453–458.

Ivanova AA (2022). The role of language in broader human cognition: evidence from neuroscience. Ch. 5. Doctoral dissertation, Massachusetts Institute of Technology.

Ivanova AA, Mineroff Z, Zimmerer V, Kanwisher N, Varley R, Fedorenko E (2021). The Language Network Is Recruited but Not Required for Nonverbal Event Semantics. Neurobiol Lang 2:176–201.

Ivanova AA, Srikant S, Sueoka Y, Kean HH, Dhamala R, O'Reilly UM, Bers MU, Fedorenko E (2020). Comprehension of computer code relies primarily on domain-general executive brain regions. elife 9:1–24.

Jackendoff R (2007). A Parallel Architecture perspective on language processing. Brain Res 1146:2–22.

Jackendoff R, Audring J (2020). Morphology and Memory: Toward an Integrated Theory. Top Cogn Sci 12:170–196.

Jacoby N, Fedorenko E (2018) Discourse-level comprehension engages medial frontal Theory of Mind brain regions even for expository texts. Lang Cogn Neurosci 35:780–796.

Jain S, Vo VA, Mahto S, LeBel A, Turek JS, Huth AG (2020). Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. Adv Neural Inf Process Syst 2020-Decem:1–12.

Janca R, Jezdik P, Cmejla R, Tomasek M, Worrell GA, Stead M, Wagenaar J, Jefferys JGR, Krsek P, Komarek V, Jiruska P, Marusic P (2015). Detection of interictal epileptiform discharges using signal envelope distribution modelling: application to epileptic and non-epileptic intracranial recordings. Brain Topogr 28:172–183.

Keshev M, Meltzer-Asscher A (2021). Noisy is better than rare: Comprehenders compromise subject-verb agreement to form more probable linguistic structures. Cogn Psychol 124:101359.

Kuperberg GR, McGuire PK, Bullmore ET, Brammer MJ, Rabe-Hesketh S, Wright IC, Lythgoe DJ, Williams SCR, David AS (2000). Common and Distinct Neural Substrates for Pragmatic, Semantic, and Syntactic Processing of Spoken Sentences: An fMRI Study. J Cogn Neurosci 12:321–341.

Lerner Y, Honey CJ, Silbert LJ, Hasson U (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. J Neurosci 31:2906–2915.

Levy R (2008a). A Noisy-Channel Model of Human Sentence Comprehension under Uncertain Input. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 234–243, Honolulu, Hawaii. Association for Computational Linguistics.

Levy R (2008b). Expectation-based syntactic comprehension. Cognition 106(3):1126–1177.

Liu X, Vermeylen L, Wisniewski D, Brysbaert M (2020). The contribution of phonological information to visual word recognition: Evidence from Chinese phonetic radicals. Cortex 133:48–64.

Lopopolo A, Frank SL, Van Den Bosch A, Willems RM (2017). Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. PLoS One 12:e0177794.

Matchin W, Hickok G (2020). The Cortical Organization of Syntax. Cereb Cortex 30:1481–1498.

Mazoyer BM, Tzourio N, Frak V, Syrota A, Murayama N, Levrier O, Salamon G, Dehaene S, Cohen L, Mehler J (1993). The Cortical Representation of Speech. J Cogn Neurosci 5:467–479.

Mesulam MM, Wieneke C, Hurley R, Rademaker A, Thompson CK, Weintraub S, Rogalski EJ (2013). Words and objects at the tip of the left temporal lobe in primary progressive aphasia. Brain 136:601–618.

Mollica F, Piantadosi ST (2019). Humans store about 1.5 megabytes of information during language acquisition. R Soc Open Sci 6: 181393.

Monti MM, Parsons LM, Osherson DN (2012). Thought Beyond Language: Neural Dissociation of Algebra and Natural Language. Psychol Sci 23:914–922.

Mukamel R, Fried I (2011). Human Intracranial Recordings and Cognitive Neuroscience. Annu Rev Psychol 63:511–537.

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

Norman-Haignere S, Kanwisher NG, McDermott JH (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. Neuron 88(6):1281-1296.

Norman-Haignere S V., Long LK, Devinsky O, Doyle W, Irobunda I, Merricks EM, Feldstein NA, McKhann GM, Schevon CA, Flinker A, Mesgarani N (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. Nat Hum Behav 6(3):455–469.

Okada K, Hickok G (2006). Identification of lexical-phonological networks in the superior temporal sulcus using functional magnetic resonance imaging. Neuroreport 17:1293–1296.

Okada K, Matchin W, Hickok G (2017). Phonological feature repetition suppression in the left inferior frontal gyrus. J Cogn Neurosci 30:1549–1557.

Overath T, McDermott JH, Zarate JM, Poeppel D (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat Neurosci 18:903–911.

Ryskin R, Futrell R, Kiran S, Gibson E (2018). Comprehenders model the nature of noise in the environment. Cognition 181:141–150.

Ryskin R, Stearns L, Bergen L, Eddy M, Fedorenko E, Gibson E (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. Neuropsychologia 158:107855.

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

Paulesu E, Frith CD, Frackowiak RSJ (1993). The neural correlates of the verbal component of working memory. Nature 362:342–345.

Paulk AC, Kfir Y, Khanna AR, Mustroph ML, Trautmann EM, Soper DJ, Stavisky SD, Welkenhuysen M, Dutta B, Shenoy K V., Hochberg LR, Richardson RM, Williams ZM, Cash SS (2022). Large-scale neural recordings with single neuron resolution using Neuropixels probes in human cortex. Nat Neurosci 2022 252 25:252–263.

Paunov AM, Blank IA, Fedorenko E (2019). Functionally distinct language and theory of mind networks are synchronized at rest and during language comprehension. J Neurophysiol 121:1244–1265.

Price CJ, Moore CJ, Humphreys GW, Wise RJS (1997). Segregating semantic from phonological processes during reading. J Cogn Neurosci 9:727–733.

Regev M, Honey CJ, Simony E, Hasson U (2013). Selective and invariant neural responses to spoken and written narratives. J Neurosci 33:15978–15988.

Regev TI, Affourtit J, Chen X, Schipper AE, Bergen L, Mahowald K, Fedorenko E (2021). High-level language brain regions are sensitive to sub-lexical regularities. bioRxiv:2021.06.11.447786.

Rodd JM, Davis MH, Johnsrude IS (2005). The Neural Mechanisms of Speech Comprehension: fMRI studies of Semantic Ambiguity. Cereb Cortex 15:1261–1269.

Rokach L, Maimon O (2005). Clustering methods. In: The data mining and knowledge discovery handbook, pp 321–352. Boston, MA.: Springer.

Saxe R, Brett M, Kanwisher N (2006). Divide and conquer: A defense of functional localizers. Neuroimage 30:1088–1096.

Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, Wolpaw JR (2004). BCI2000: A general-purpose brain-computer interface (BCI) system. IEEE Trans Biomed Eng 51:1034–1043.

Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum JB, Fedorenko E (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proc Natl Acad Sci USA 118.

Scott TL, Gallée J, Fedorenko E (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. Cogn Neurosci 8:167–176.

Shain C, Blank IA, Fedorenko E, Gibson E, Schuler W (2022a). Robust Effects of Working Memory Demand during Naturalistic Language Comprehension in Language-Selective Cortex. J Neurosci 42:7412–7430.

Shain C, Blank IA, van Schijndel M, Schuler W, Fedorenko E (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. Neuropsychologia 138:107307.

Shain C, Kean H, Lipkin B, Affourtit J, Siegelman M, Mollica F, Fedorenko E, authors C (2021). 'Constituent length' effects in fMRI do not provide evidence for abstract syntactic processing. bioRxiv:2021.11.12.467812.

Shain C, Paunov A, Chen X, Lipkin B, Fedorenko E (2022b). No evidence of theory of mind reasoning in the human language network. bioRxiv:2022.07.18.500516.

Toneva M, Wehbe L (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). Adv Neural Inf Process Syst 2019-Decem:14954–14964.

Vagharchakian L, Dehaene-Lambertz G, Pallier C, Dehaene S (2012). A temporal bottleneck in the language comprehension network. J Neurosci 32:9089–9102.

Vandenberghe R, Nobre AC, Price CJ (2002). The Response of Left Temporal Cortex to Sentences. J Cogn Neurosci 14:550–560.

Supplementary Information

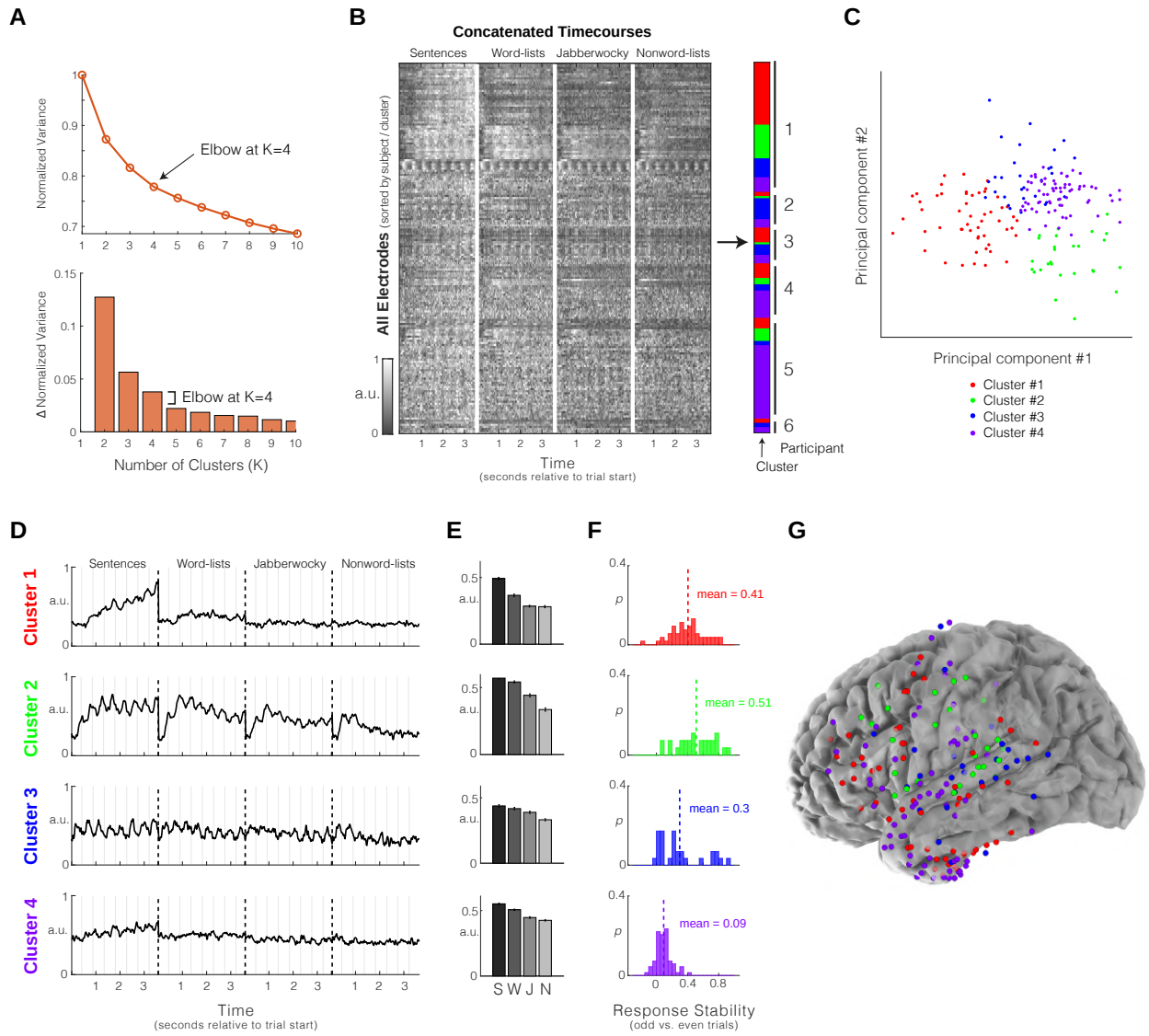

Figure S1 – Dataset 1 k-means clustering results. A) Search for optimal k using the "elbow method". Top: variance (sum of the distances of all electrodes to their assigned cluster center) normalized by the variance when k=1 as a function of k (normalized variance (NV)). Bottom: change in NV as a function of k (NV(k+1) – NV(k)). After k=4 the change in variance becomes more moderate, suggesting that 4 clusters appropriately describe this dataset. **B)** Clustering mean electrode responses (S+W+J+N) using k-means (k=4) with squared-Euclidean distance. Shading of the data matrix reflects normalized high-gamma power (70-150Hz). **C)** Electrode responses visualized on their first two principal components, colored by cluster. **D)** Average timecourse by cluster. Clusters 1-3 resemble the clusters reported in **Figure 3**, and Cluster 4 is qualitatively similar to Cluster 1 with a less pronounced increase of neural activity over the course of a sentence. **E)** Mean condition responses by cluster. Error bars reflect standard error. **F)** Response reliability across odd and even trials (based on a correlation of mean responses) by cluster. Electrodes in the Cluster 4 (the cluster that does not emerge when k-medoids clustering is used) display the least reliable responses to linguistic stimuli relative to the other clusters. **G)** Anatomical distribution of clusters across all participants (n=6).

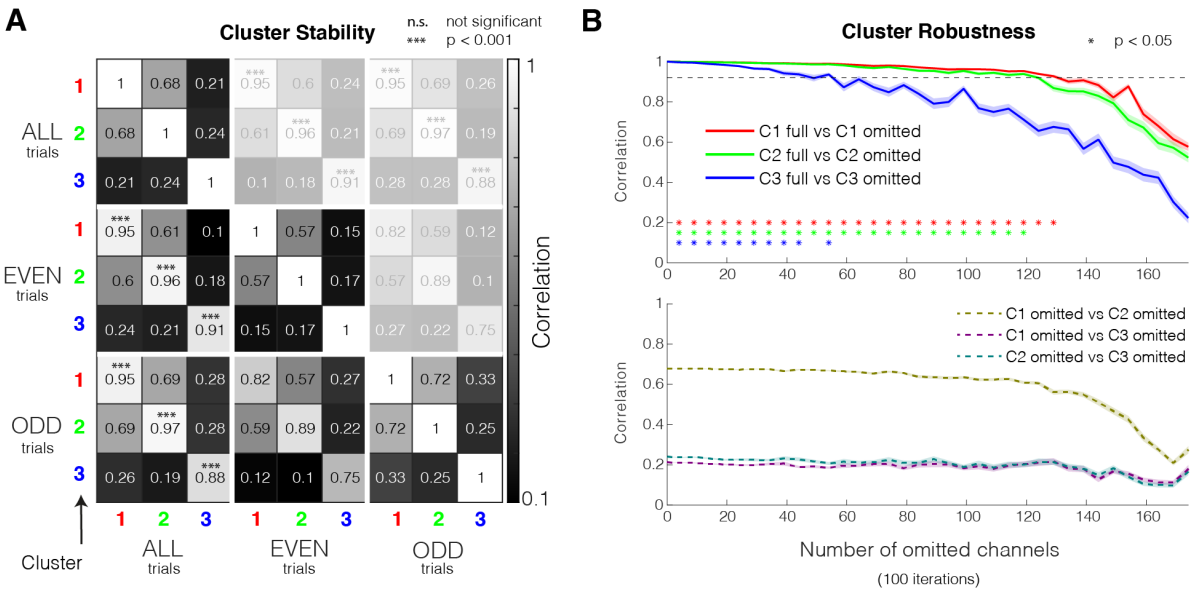

Figure S2 – Cluster evaluation within Dataset 1. A) Comparison of clusters from all trials (top three rows) versus only even (middle three rows) or odd (bottom three rows) trials. Clusters that emerge using only odd or even trials are strikingly similar to the clusters that emerge when all trials are used (ps<0.001; evaluated against clustering solutions when trials are shuffled; Methods). **B)** Robustness of clusters electrode omission (random subsets of electrodes were removed in increments of 5). Top: Similarity of cluster centers when all electrodes were used versus when subsets of electrodes were removed. Stars reflect significant similarity with the full dataset (p<0.05; evaluated against clustering solution when trials are shuffled; Methods). Shaded regions reflect standard error. Cluster 3 is driven the most by individual electrodes relative to Clusters 1 and 2. Bottom: Similarity between cluster centers when subsets of electrodes are removed. Shaded regions reflect standard error. The relationship between clusters (e.g., similarity of Clusters 1 and 2) remains stable (very little change in correlation) until the cluster centers no longer resemble the cluster centers from all electrodes (see Top).

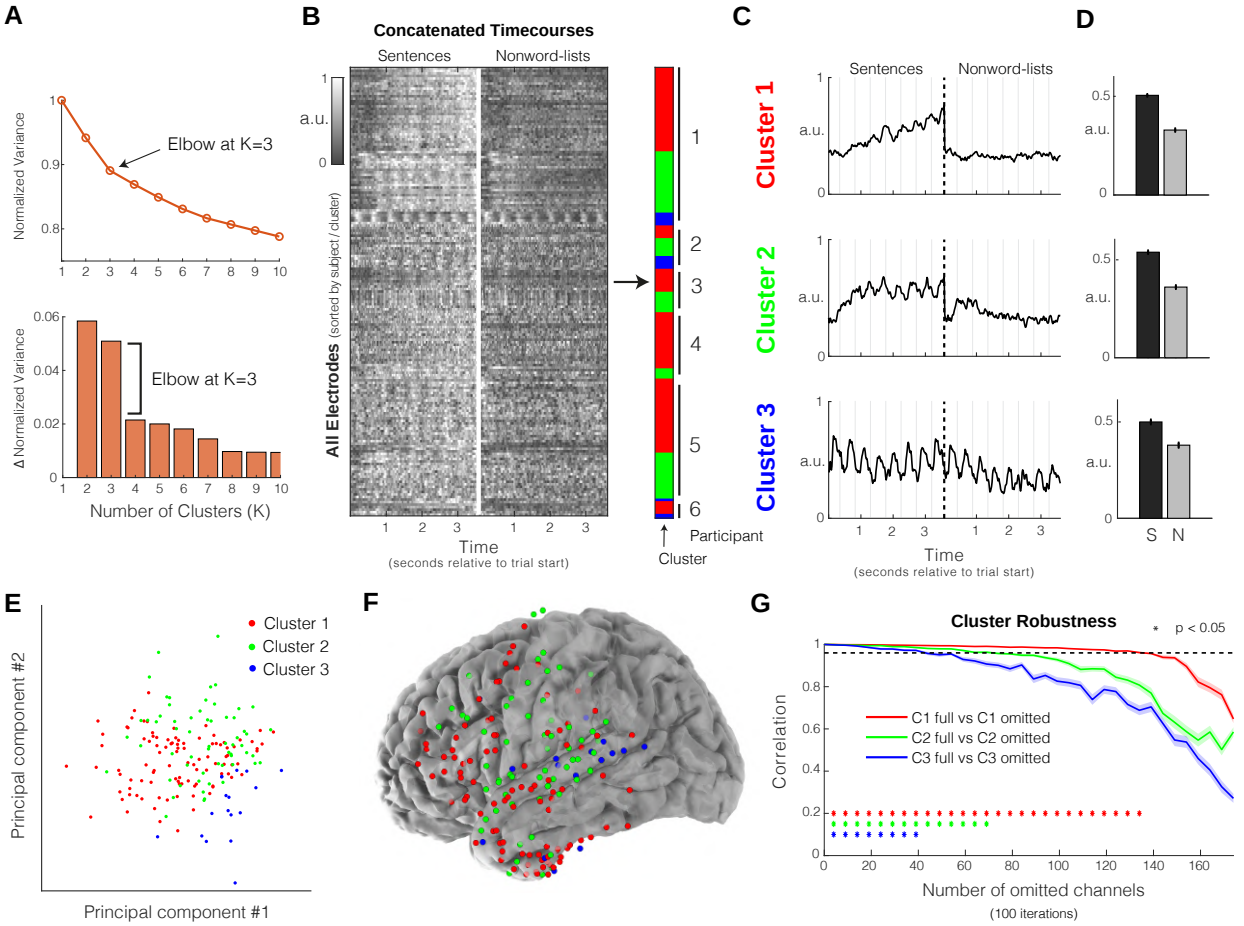

Figure S3 – Dataset 1 k-medoids clustering results with only S-N conditions. A) Search for optimal k using the "elbow method". Top: variance (sum of the distances of all electrodes to their assigned cluster center) normalized by the variance when k=1 as a function of k (normalized variance (NV)). Bottom: change in NV as a function of k (NV(k+1) – NV(k)). After k=3 the change in variance becomes more moderate, suggesting that 3 clusters appropriately describe Dataset 1 when using only the responses to sentences and nonwords (as was the case when all four conditions were used). **B)** Clustering mean electrode responses (only S+N, importantly) using k-medoids (k=3) with a correlation-based distance. Shading of the data matrix reflects normalized high-gamma power (70-150Hz). **C)** Average timecourse by cluster. Clusters 1-3 show striking similarity to the clusters reported in **Figure 3**. **D)** Mean condition responses by cluster. Error bars reflect standard error. **E)** Electrode responses visualized on their first two principal components, colored by cluster. **F)** Anatomical distribution of clusters across all participants (n=6). **G)** Robustness of clusters electrode omission (random subsets of electrodes were removed in increments of 5). Stars reflect significant similarity with the full dataset ($p<0.05$; evaluated against clustering solution when trials are shuffled; Methods). Shaded regions reflect standard error. Relative to when all conditions were used, Cluster 2 is less robust to electrode omission (although still more robust than Cluster 3), suggesting that responses to word-lists and Jabberwocky sentences (both not present here) are particularly important for distinguishing Cluster 2 electrodes from Cluster 1 and 3 electrodes.

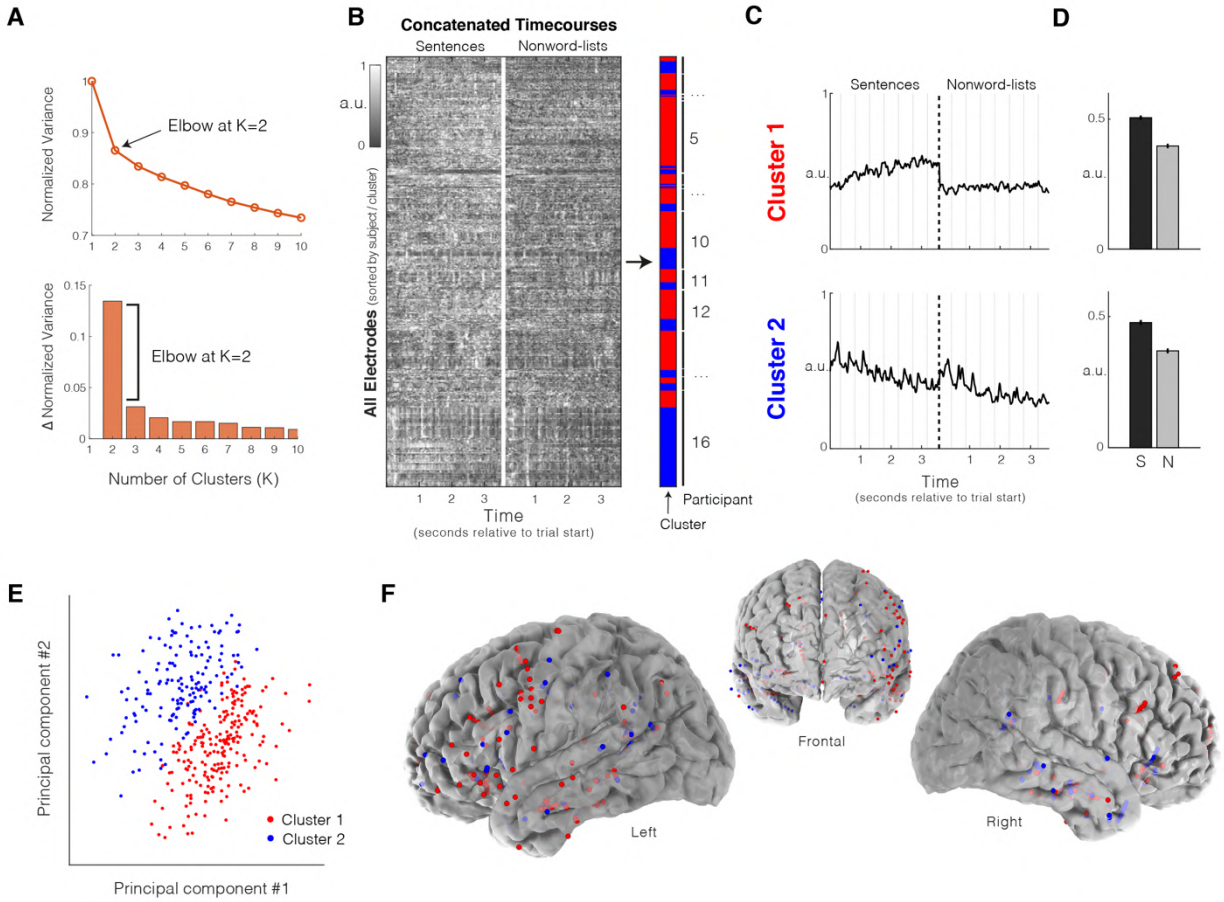

Figure S4 – Dataset 2 k-medoids clustering results with k=2. A) Search for optimal k using the "elbow method". Top: variance (sum of the distances of all electrodes to their assigned cluster center) normalized by the variance when k=1 as a function of k (normalized variance (NV)). Bottom: change in NV as a function of k (NV(k+1) – NV(k)). After k=2 the change in variance becomes more moderate, suggesting that 2 clusters appropriately describe Dataset 2 (for a direct comparison of Dataset 1 and Dataset 2 when k=3, see **Figure 7**). **B)** Clustering mean electrode responses (S+N) using k-medoids (k=2) with a correlation-based distance. Shading of the data matrix reflects normalized high-gamma power (70-150Hz). **C)** Average timecourse by cluster. Here, Clusters 1 and 2 are significantly similar to Clusters 1 and 3, respectively, from **Figure 3** (ps<0.001, evaluated against clustering solutions when trials are shuffled; Methods). **D)** Mean condition responses by cluster. Error bars reflect standard error. **E)** Electrode responses visualized on their first two principal components, colored by cluster. **F)** Anatomical distribution of clusters across all participants (n=16).

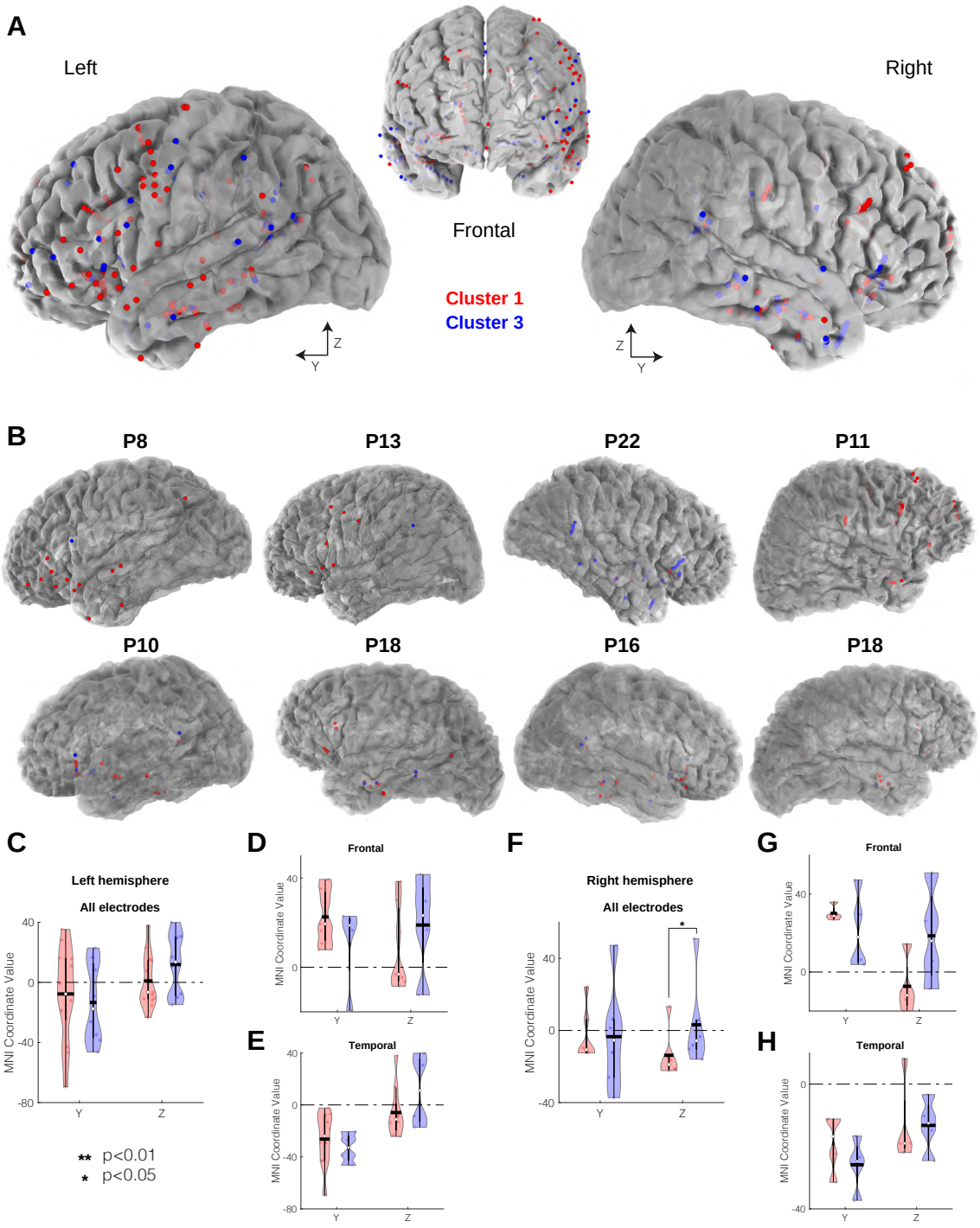

Figure S5 – Anatomical distribution of the clusters in Dataset 2. Anatomical distribution of language-responsive electrodes in Dataset 2 across all subjects in MNI space, colored by cluster. Only Clusters 1 and 3 (those from Dataset 1 that replicate to Dataset 2) are shown. **B)** Anatomical distribution of language-responsive electrodes in subject-specific space for eight sample subjects. **C-H)** Violin plots of MNI coordinate values for Clusters 1 and 3 in the left and right hemisphere (**C-E** and **F-H**, respectively), where plotted points represent the mean of all coordinate values for a given participant and cluster. The mean is plotted with a black horizontal line, and the median is shown with a white circle. Significance values are computed using a linear mixed-effects model (LME, see **Tables S3** and **S4**; Methods). The Cluster 3 posterior bias from Dataset 1 is weakly present but not statistically reliable.

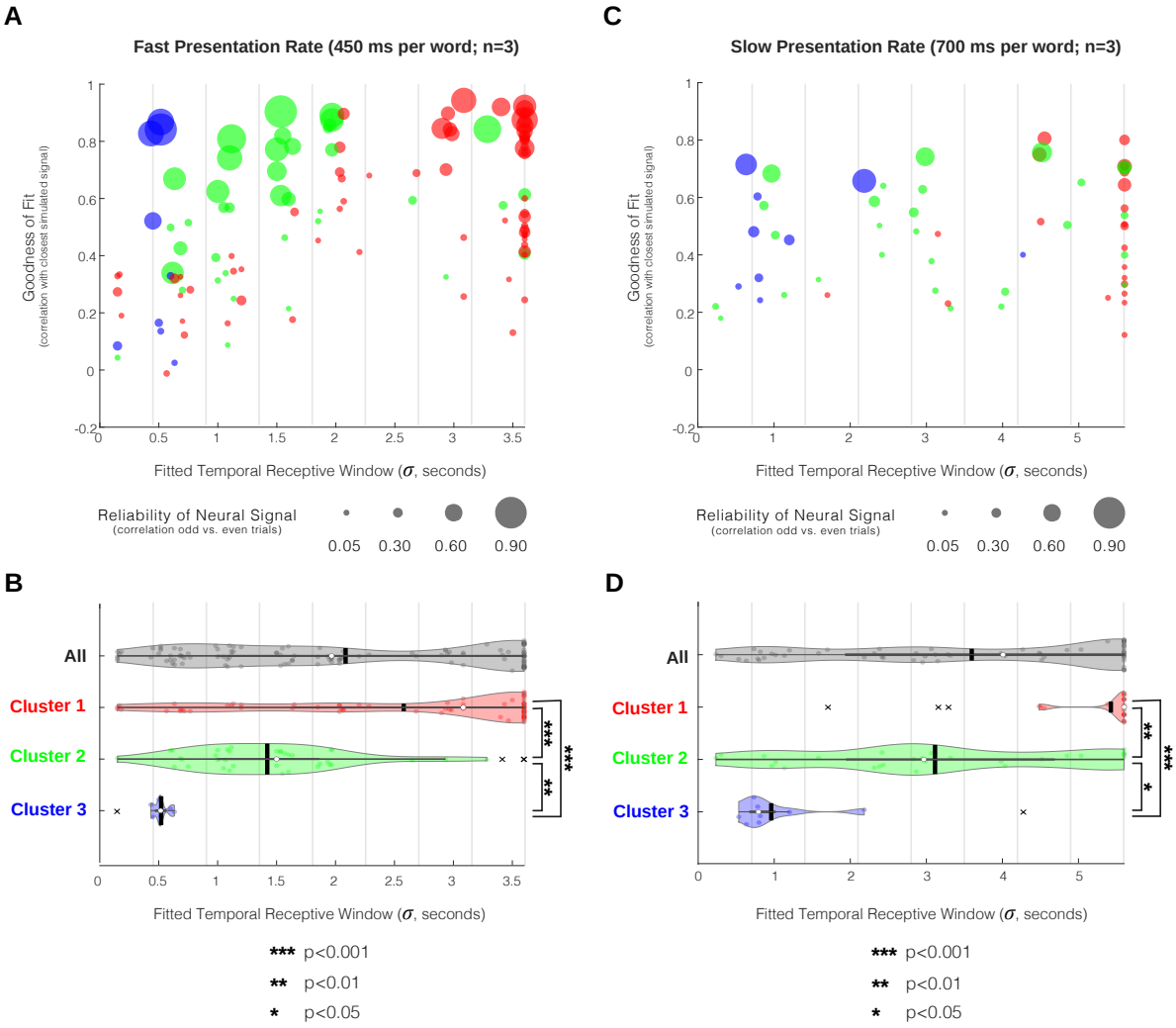

Figure S6 – Estimation of temporal receptive window sizes for electrodes in Dataset 1, separated by presentation rate. As in **Figure 5** but separating the participants by slow or fast presentation rate. **A,C)** Best TRW fit (using a model that simulates neural responses to the sentence condition as a convolution of a simplified stimulus train and gaussian kernels with varying widths) for all electrodes colored by cluster and sized by the reliability of the neural signal as estimated by correlating responses to odd and even trials (**Figure 2D**) . The 'goodness of fit', or correlation between the simulated and observed neural signal (sentence condition only), is shown on the y-axis. **A)** Best TRW fit for the 3 participants that saw the fast presentation rate (450 ms per word/nonword). **C)** Best TRW fit for the 3 participants that saw the slow presentation rate (700 ms per word/nonword). **B,D)** Estimated TRW sizes across all electrodes (grey) and per cluster (red, green, and blue). Black vertical lines correspond to the mean window size and the white dots correspond to the median. "x" marks indicate outliers (more than 1.5 interquartile ranges above the upper quartile or less than 1.5 interquartile ranges below the lower quartile). Significance values are calculated using a linear mixed-effects model (LME, Methods). **B)** Estimated TRW sizes for the 3 participants that saw the fast presentation rate (450 ms per word/nonword). **D)** Estimated TRW sizes for the 3 participants that saw the slow presentation rate (700 ms per word/nonword). The similarity of the TRW distributions across the two presentation rates suggest that the TRW of these electrodes are language-, not time-dependent.

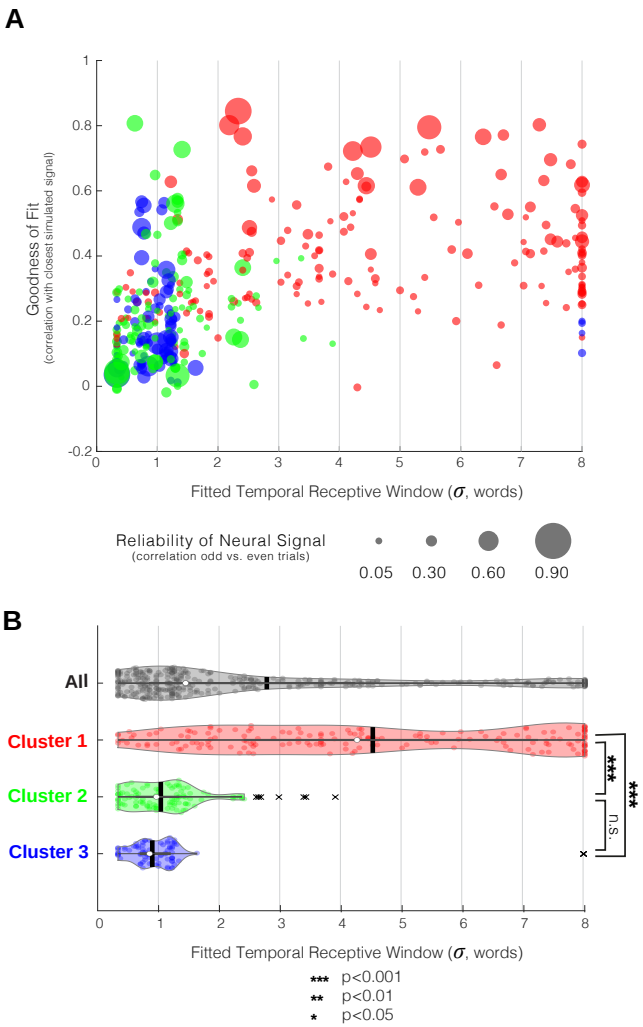

Figure S7 – Estimation of temporal receptive window sizes for electrodes in Dataset 2. As in **Figure 5** but for electrodes in Dataset 2. **A)** Best TRW fit (using a model that simulates neural responses to the sentence condition as a convolution of a simplified stimulus train and gaussian kernels with varying widths) for all electrodes colored by cluster and sized by the reliability of the neural signal as estimated by correlating responses to odd and even trials (**Figure 2D**). The 'goodness of fit', or correlation between the simulated and observed neural signal (sentence condition only), is shown on the y-axis. **B)** Estimated TRW sizes across all electrodes (grey) and per cluster (red, green, and blue). Black vertical lines correspond to the mean window size and the white dots correspond to the median. "x" marks indicate outliers (more than 1.5 interquartile ranges above the upper quartile or less than 1.5 interquartile ranges below the lower quartile). Significance values are calculated using a linear mixed-effects model (LME, Methods).

Name	Estimate	SE	tStat	DF	pValue
Intercept: Cluster 1, Condition S	0.102	0.022	4.651	696	4.0E-06
Cluster 2	0.077	0.035	2.180	696	3.0E-02
Cluster 3	0.241	0.074	3.237	696	1.3E-03
Condition W	-0.023	0.025	-0.885	696	3.8E-01
Condition J	-0.036	0.027	-1.344	696	1.8E-01
Condition N	-0.061	0.026	-2.317	696	2.1E-02
coefTest Cluster 3 vs. 2	NaN	NaN	2.14	1	0.03
coefTest Condition N vs. W	NaN	NaN	1.45	1	0.15
coefTest Condition J vs. W	NaN	NaN	0.50	1	0.62
coefTest Condition N vs. J	NaN	NaN	0.85	1	0.40
Cluster 2 * Condition W	-0.03	0.04	-0.66	696	0.51
Cluster 3 * Condition W	0.02	0.06	0.35	696	0.73
Cluster 2 * Condition J	-0.04	0.04	-1.06	696	0.29
Cluster 3 * Condition J	0.01	0.06	0.10	696	0.92
Cluster 2 * Condition N	-0.06	0.04	-1.60	696	0.11
Cluster 3 * Condition N	-0.04	0.06	-0.71	696	0.48

Table S1 LME results quantifying degree of stimulus locking by cluster. All estimates from the linear mixed-effects model (LME) regressing the locking value (Methods) on the categorical variables of cluster (3 levels) and condition (4 levels for sentences (S), word-lists (W), jabberwocky (J), nonword-lists (N), Methods), including their interaction, all grouped by the random variable of participant. Model formula: *Locking ~ cluster*condition + (cluster|participant) + (condition|participant)*). Implemented with Matlab *fitlme* routine. Asterisks represent interactions. The first level of the variables was modeled as an intercept (Cluster 1, Condition: S) and all other estimates were evaluated and compared statistically to the intercept. In order to compare other pairs of estimates we ran a coefficient test post-hoc using Matlab routine *coeftest*. The pairwise comparisons of all 3 clusters were significant (all ps<0.05). The only pairwise comparison between conditions that was significant was S vs. N ($p<0.05$), and all other condition comparisons did not reach significance. All interaction terms were not significant. An additional ANOVA test for LME revealed a significant main effect for cluster ($F(2,696)=5.6$, $p<0.01$) and the main effect for condition as well as the interaction term did not reach significance. See **Figure 5**.

Dataset	Coordinate	Name	Estimate	SE	tStat	DF	pValue
Dataset 1	y	Cluster 1	1.38	3.77	0.37	171	0.71421
Dataset 1	y	Cluster 2	-2.93	3.57	-0.82	171	0.41348
Dataset 1	y	Cluster 3	-21.15	5.87	-3.60	171	0.00041
Dataset 1	y	coefTest Cluster3-2	NaN	NaN	-3.11	1	0.00219
Dataset 1	z	Cluster 1	-4.82	4.38	-1.10	171	0.27330
Dataset 1	z	Cluster 2	17.37	6.94	2.50	171	0.01332
Dataset 1	z	Cluster 3	11.08	8.19	1.35	171	0.17825
Dataset 1	z	coefTest Cluster3-2	NaN	NaN	-0.82	1	0.41411

Table S2 – LME results comparing MNI coordinates of the 3 clusters, Dataset 1, Left hemisphere.
All estimates from the linear mixed-effects model (LME) regressing the y (posterior-anterior) and z (inferior-superior) MNI coordinates (Methods) on the categorical variable of cluster (3 levels) grouped by the random variable of participant. Model formula: *MNI coordinate ~ cluster + (cluster|participant).* Details are similar to Table S1. The y-coordinate of Cluster 3 was significantly different from Clusters 1 and 2 (ps<0.01). All the other comparisons did not reach significance. See **Figure 6**.

Dataset	Coordinate	Name	Estimate	SE	tStat	DF	pValue
Dataset 2	y	Cluster 1	-5.20	7.08	-0.73	196	0.4636
Dataset 2	y	Cluster 2	-7.28	9.80	-0.74	196	0.4584
Dataset 2	y	Cluster 3	-0.97	10.34	-0.09	196	0.9255
Dataset 2	y	coefTest_Cluster3-2	NaN	NaN	0.81	1	0.4210
Dataset 2	z	Cluster 1	3.10	4.03	0.77	196	0.4426
Dataset 2	z	Cluster 2	12.71	3.82	3.32	196	0.0011
Dataset 2	z	Cluster 3	7.20	5.84	1.23	196	0.2185
Dataset 2	z	coefTest_Cluster3-2	NaN	NaN	0.95	1	0.3451

Table S3 – LME results comparing coordinates of the 3 clusters, Dataset 2, Left hemisphere.
Similar to Table S1 but for Dataset 2, left hemisphere electrodes. The only significant comparison was the z-coordinate of Cluster 2 relative to Clusters 1 ($p<0.01$). See **Figure S5**.

Dataset	Coordinate	Name	Estimate	SE	tStat	DF	pValue
Dataset 2	y	Cluster 1	2.5	6.9	0.4	160	0.717
Dataset 2	y	Cluster 2	-17.3	6.9	-2.5	160	0.014
Dataset 2	y	Cluster 3	-7.5	7.1	-1.1	160	0.294
Dataset 2	y	coefTest_Cluster3-2	NaN	NaN	1.4	1	0.178
Dataset 2	z	Cluster 1	-8.4	5.7	-1.5	160	0.143
Dataset 2	z	Cluster 2	5.1	4.7	1.1	160	0.271
Dataset 2	z	Cluster 3	14.5	5.7	2.5	160	0.012
Dataset 2	z	coefTest_Cluster3-2	NaN	NaN	2.2	1	0.028

Table S4 – LME results comparing coordinates of the 3 clusters, Dataset 2, Right hemisphere.
Similar to Table S3 but for right-hemisphere electrodes. The significant comparisons were the y-coordinates of Cluster 2 vs. 1 and the z-coordinates of Cluster 3 relative to clusters 2 and 1 ($p<0.05$). See **Figure S5**.

Name	Estimate	SE	tStat	DF	pValue
Cluster 1	6.5	0.5	13.9	174	6.9E-30
Cluster 2	-2.5	0.6	-4.5	174	1.3E-05
Cluster 3	-5.0	0.6	-8.2	174	4.6E-14
coefTest_Cluster3-2	NaN	NaN	4.14	1	5.4E-05

Table S5 – LME results comparing temporal receptive windows (TRW) of the 3 clusters, Dataset 1.
All estimates from the linear mixed-effects model (LME) regressing the estimated temporal receptive window (trw) size (Methods) on the categorical variable of cluster (3 levels) grouped by the random variable of participant. Model formula: *trw ~ cluster + (cluster|participant).* Details are similar to Table S1. All comparisons were statisitically significant: Cluster 2 had a smaller trw compared to Cluster 1, and Cluster 3 had the smallest trw compared to both other clusters (all ps<0.0001). See **Figure 4**.

Name	Estimate	SE	tStat	DF	pValue
Cluster 1	2.6	0.3	10.3	114	6.4E-18
Cluster 2	-1.0	0.2	-4.3	114	3.6E-05
Cluster 3	-2.1	0.4	-4.9	114	3.8E-06
coefTest_Cluster3-2	NaN	NaN	2.8	1	5.9E-03

Table S6 – LME results comparing temporal receptive windows (TRW) of the 3 clusters, Dataset 1, only participants with the faster presentation rate (450 ms, n=3).
Similar to Table S5, but only participants with the faster presentation rate (450 ms, n=3). All comparisons are significant (ps<0.0001). See **Figure S6**.

Name	Estimate	SE	tStat	DF	pValue
Cluster 1	5.2	0.3	15.4	57	6.4E-22
Cluster 2	-2.3	0.8	-2.9	57	5.2E-03
Cluster 3	-3.7	0.7	-5.4	57	1.2E-06
coefTest_Cluster3-2	NaN	NaN	2.62	1	0.011

Table S7 – LME results comparing temporal receptive windows (TRW) of the 3 clusters, Dataset 1, only participants with the slower presentation rate (700 ms, n=3).
Similar to Table S6, but only participants with the slower presentation rate (700 ms, n=3). All comparisons are significant (ps<0.05). See **Figure S6**.

Name	Estimate	SE	tStat	DF	pValue
Cluster 1	4.5	0.3	14.2	359	1.8E-36
Cluster 2	-3.3	0.4	-8.9	359	3.4E-17
Cluster 3	-3.2	0.3	-9.8	359	3.1E-20
coefTest_Cluster3-2	NaN	NaN	0.12	1	0.91

Table S8 – LME results comparing temporal receptive windows (TRW) of the 3 clusters, Dataset 2, using 8 words.
Similar to Table S5, but for Dataset 2 using the first 8 words per each trial. All comparisons except for Cluster 3 vs. 2 are significant (ps<0.0001). See **Figure S7**.

Name	Estimate	SE	tStat	DF	pValue
Cluster 1	5.3	0.4	12.7	359	1.2E-30
Cluster 2	-3.9	0.5	-8.1	359	1.1E-14
Cluster 3	-2.6	0.7	-4.0	359	9.1E-05
coefTest_Cluster3-2	NaN	NaN	1.6	1	0.11

Table S9 – LME results comparing temporal receptive windows (TRW) of the 3 clusters, Dataset 2, using 12 words.

Similar to Table S8, but for Dataset 2 using the full 12 words in a trial. All comparisons except for Cluster 3 vs. 2 are significant (ps<0.0001).