Intracranial recordings reveal three distinct neural response patterns in the language network

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

[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]Albany Medical Center, Albany NY
[6]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

Abstract

A left-lateralized network of frontal and temporal brain regions is specialized for language processing—spoken, written, or signed. Different regions of this 'language network' have all been shown to be sensitive to various forms of linguistic information, from combinatorial sentence structure to word meanings, to sub-lexical regularities. However, whether neural computations are the same across and within these different brain regions remains debated. Here, we examine responses during language processing recorded intracranially in patients with intractable epilepsy. Across two datasets (Dataset 1: n=6 participants, m=177 language-responsive electrodes; Dataset 2: n=16 participants, m=362 language-responsive electrodes), we clustered language-responsive electrodes and found three distinct response profiles, with differences in response magnitude between linguistic conditions (e.g., sentences vs. lists of words), different temporal dynamics over the course of the stimulus, and different degrees of stimulus locking. We argue that these profiles correspond to different temporal receptive windows that vary in size between sub-lexical units and multi-word sequences. These results demonstrate the functional heterogeneity of neural responses in the language network and highlight the diversity of neural computations that may be needed in order to extract meaning from linguistic input. Importantly, electrodes that exhibit these distinct profiles do not cluster spatially and are instead interleaved across frontal and temporal language areas, which likely made It difficult to uncover functional differences in past fMRI studies. This mosaic of neural responses across the language network suggests that all language regions have direct access to distinct response types—a property that may be crucial for the efficiency and robustness of language processing mechanisms.

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 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., 2021). 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 (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, with distinct temporal profiles. We speculate that these response types relate to the timescales of information integration in the language system (Lerner et al., 2011; Blank and Fedorenko, 2020).

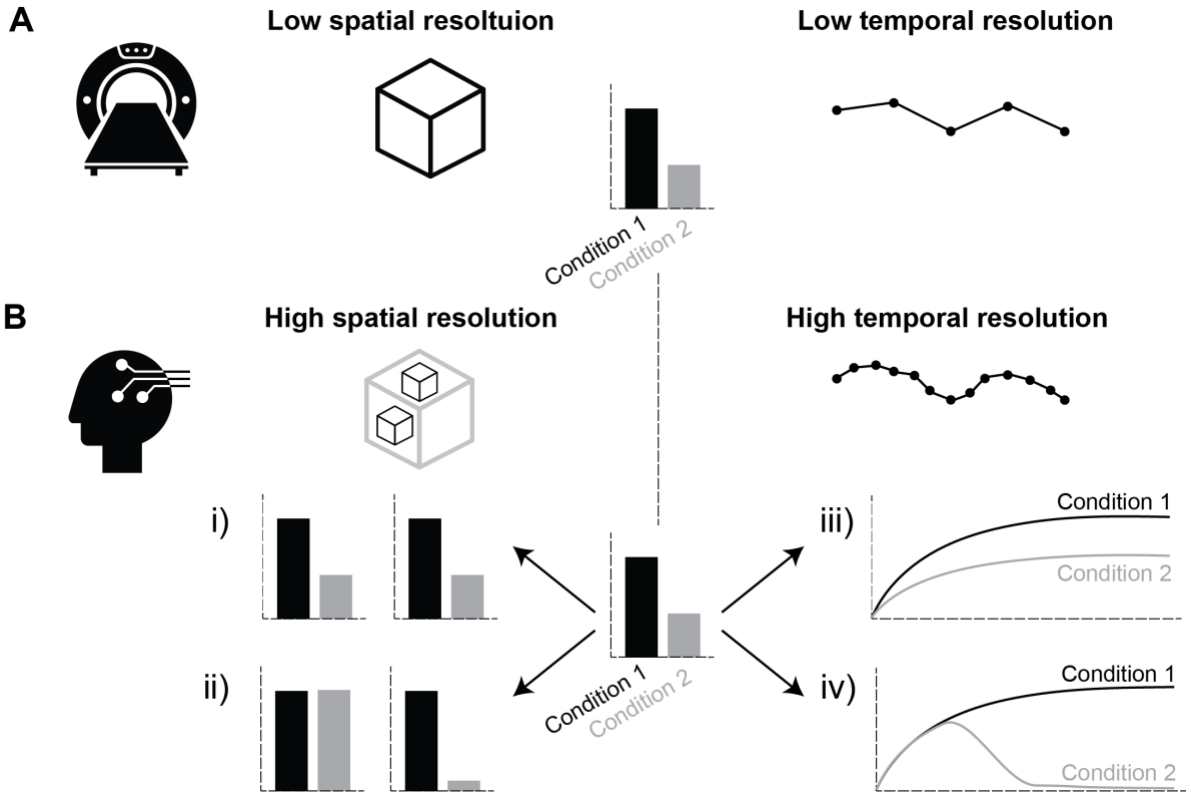

Figure 1 – **Low spatial and temporal resolution may obscure details of neural computations**. **A)** Hypothetical effect sizes for 2 experimental conditions as measured using fMRI, with relatively low spatial and temporal resolution. **B)** Illustration of the same hypothetical experimental conditions 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).

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 channels (see Language-Responsive Channel 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). Channel 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 (channels; 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 channels, trigger channels, scalp electroencephalography (EEG) channels, EKG channels), as well as d) those with significant line noise, defined as channels with line noise greater than 5 standard deviations above other channels, 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—channels 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' channels fit this criterion.) These exclusion criteria left 108, 115, 92, 106, 93, and 36 channels for analysis for the 6 participants in Dataset 1 (**Table 1**) and between 76 and 228 channels for the 16 participants in Dataset 2 (**Table 2**).

Next, the common average reference was removed from all channels connected to the same amplifier. 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 90th 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 channels, 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 resampled to the faster version (words presented for 450 ms each).

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.

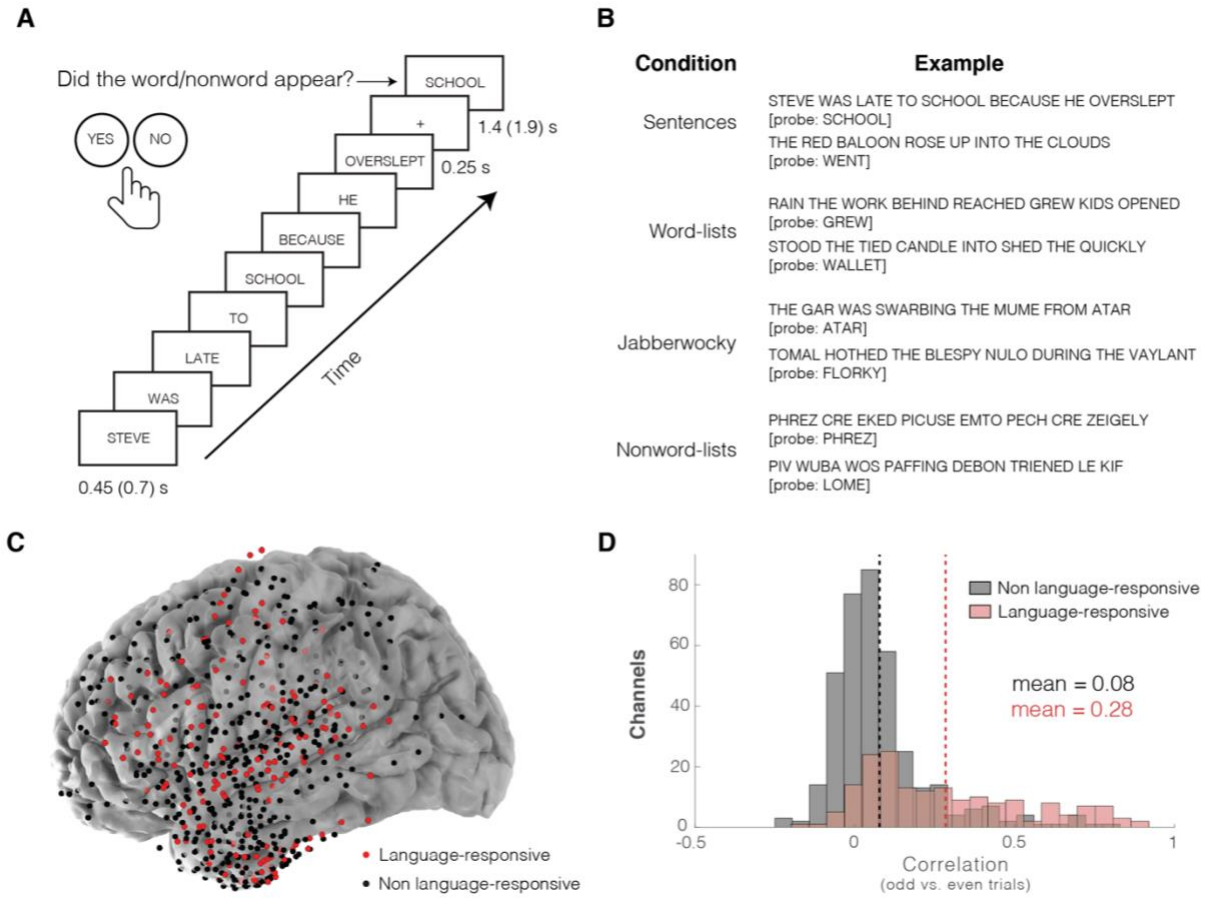

Figure 2 – Experimental procedure and the distribution of the implanted electrodes for Dataset 1. A) A sample trial from the sentences condition. **B)** Two sample items from each of the four experimental conditions. **C)** The locations of language-responsive (n=177, red; see Language-Responsive Channel 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 channels found across the cortex. **D)** Response stability across odd and even trials for language-responsive and non language-responsive channels. Language-responsive channels exhibit more reliable responses to linguistic stimuli than non language-responsive channels.

Language-Responsive Channel Selection

In both datasets, we identified language-responsive channels as channels 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 channel (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 channel 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. Channels 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 channels to be included in the critical analyses, and to increase the chances of discovering distinct response profiles. The stability/robustness analyses and generalizing the critical results across two independent datasets minimize the possibility of false positive / accidental findings.) The majority of the language-responsive channels (98.3% in Dataset 1, 53.9% in Dataset 2) fell in the left hemisphere, but we use channels 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 channels (S>N)	Total clean channels	Total channels	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 channels' excludes reference channels, ground channels, microphone channels, trigger channels, skull EEG channels, and EKG channels; and 'Total clean channels' excludes channels 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 channels (S>N)	Total clean channels	Total channels	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)).

Analyses

We asked two research questions: 1) whether the language network contains meaningfully different response profiles, and 2) whether channels with different response profiles exhibit anatomical topographic biases.

1) Does the language network contain meaningfully different response profiles?

We asked whether language-responsive electrodes differ in their temporal response profiles. This analysis proceeded in three steps: i) cluster discovery and characterization using Dataset 1, ii) cluster evaluation, and iii) an attempt to replicate the clusters using Dataset 2.

i) Cluster discovery and characterization using Dataset 1

Using Dataset 1 (n=6 participants, m=177 language-responsive channels), we created a single timecourse per channel 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—channels in our case—into k groups, where k is predetermined. The algorithm attempts to minimize the distances between each channel and the cluster center, where cluster centers are represented by 'medoids' (exemplar channels) 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 which evaluates the similarity between channels by the extent to which

their timecourses covary rather than how far apart they are in high-dimensional Euclidean space. In pilot analyses, we found that the Euclidean distance metric, commonly used in the k-means approach, could exaggerate distances between covarying channels (but see **Figure S1** for evidence that similar clusters emerge with a k-means clustering algorithm).

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 channels 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.

To quantitatively characterize the clusters, we performed two analyses, which focus on two observable differences in the response profile. The first analysis examined the *timecourse of condition divergence*, as quantified by the channels' 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 channel, 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 channels 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.

The second analysis examined the degree of word-by-word locking to the stimulus. To estimate the degree of stimulus locking for each channel 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 channels and conditions on the fixed effects categorical variable of *cluster* (with 3 levels for Cluster 1, 2 or 3 according to which cluster each channel 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.

ii) Cluster evaluation within Dataset 1

We performed two analyses to evaluate the stability and robustness of the clusters within Dataset 1.

1) Stability of clusters across trials: We evaluated the stability of the clustering solution by performing the same clustering procedure as described in i 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 channels 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 channels 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 channel should be interchangeable with responses in another channel. Using half of the data, we shuffled individual trials across channels (within each condition separately), re-clustered the channels 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).)

In a complementary approach, we also evaluated the proportions of channels that get assigned to the same cluster when clustering based on all versus half of the data (with cluster correspondence determined by matching cluster centers across solutions, as described above).

2) Robustness of clusters to data loss: We evaluated the robustness of the clustering solution in i to loss of channels to ensure that the solution was not driven by particular channels or participants.

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

To statistically evaluate whether the clustering solutions with only a subset of channels were more similar to the solution on the full set of channels 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 channels (within each condition separately), re-clustered the channels 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 channels and a solution on a subset of the channels) 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 channels, 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 channels was no longer more correlated to the solution on the full set of channels than would be expected by chance.

iii) 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 in i 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 channels). 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 in ii-1. In particular, using Dataset 2, we shuffled individual trials across channels (within each condition separately), re-clustered the channels 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).

2) Do channels with different response profiles exhibit anatomical topographic biases?

We examined the anatomical topographic distribution of the channels 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., channels in Cluster 1 vs. 2) with respect to the three axes: left-right [x], anterior-posterior [y], and inferior-superior [z]. This approach allowed us to ask whether, for example, channels that belong to one cluster tend to consistently fall posterior to the channels that belong to another cluster.

To do this, we extracted the MNI coordinates of all the channels 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 channel 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 channels in Cluster 2 to those in Cluster 1, and channels in Cluster 3 to those in Cluster 1. To additionally compare channels 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 channels in the two hemispheres separately.

Results

1. Language-responsive channels exhibit three distinct response profiles.

We clustered the gamma-band neural response patterns of language-responsive channels from Dataset 1 (6 participants, 177 language-responsive channels; **Methods**, **Table 1**) to sentences (S), word lists (W), Jabberwocky sentences (J) and nonword lists (N) (**Figure 2**). The k-medoids clustering algorithm, combined with the "elbow" method (Methods), suggested that three clusters (k=3) optimally explained the data (**Figure 3A**). Although we combined the channels from all 6 participants for clustering, channels that belong to Cluster 1 (92 total channels) and Cluster 2 (67 total channels) were evident in every participant individually, and channels that belong to Cluster 3 (18 total channels) were evident in 4 of the 6 participants (**Figure 3B**).

Inspection of the average timecourse by cluster (**Figure 3E**) revealed three distinct response profiles (see **Figure 3D** for representative channels (medoids) from each cluster). Cluster 1 (n=92 channels) 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 channels) displayed a quicker build-up of neural activity in the S condition that plateaued approximately 3 words into the sentence, a quick buildup 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 stimulus (each word's onset) in the S condition. Finally, Cluster 3 (n=18 channels) 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. The magnitudes of response were also more similar across conditions for Cluster 3, although the S >W>J>N pattern was still present (**Figure 3F**).

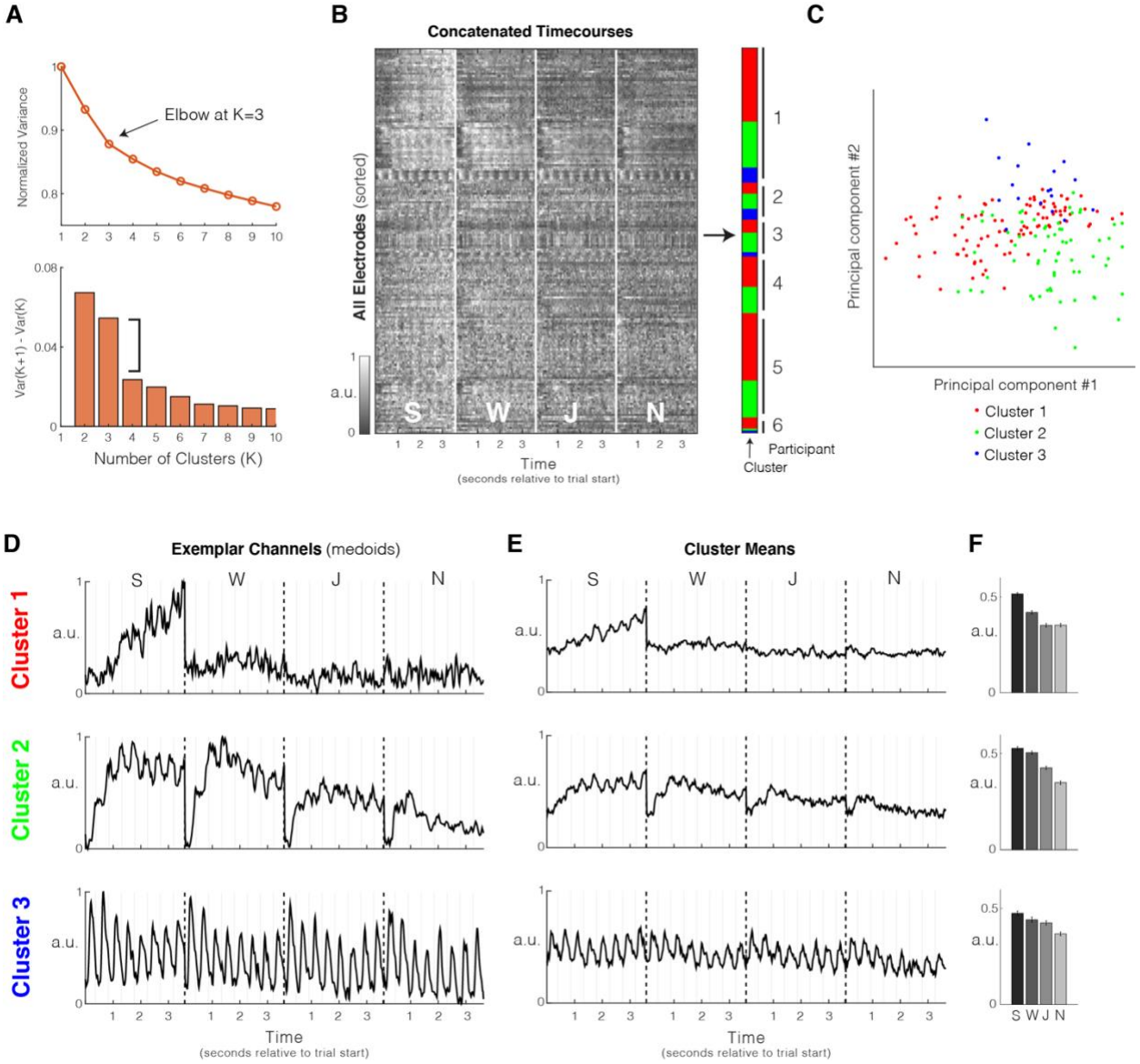

Figure 3 – Dataset 1 k-medoids clustering results. A) Search for optimal k using the "elbow method". Top: variance normalized by the variance when k=1 as a function of k. Bottom: change in normalized variance as a function of k. After k=3 the change in variance becomes more moderate, suggesting that 3 clusters appropriately describe this dataset. **B)** Clustering mean channel responses (S+W+J+N) using k-medoids (k=3) with a correlation-based distance. Shading of the data matrix reflects normalized high-gamma power. Clusters 1 and 2 (red and green, respectively) are present in all six participants, and Cluster 3 (blue) is present in four of six particiants. **C)** Channel responses visualized on their first two principal components, colored by cluster. **D)** Timecourses of representative channels ('medoids') from each of the three clusters. Here and in E, vertical gray lines represent word/nonword onsets. **E)** Timecourses averaged across all channels in each cluster. Clusters exhibit distinct response profiles (quantitatively described in **Figure 4**). **F)** Mean condition responses by cluster. After averaging across time, response profiles are not as distinct by cluster, underscoring the importance of temporal information in elucidating this grouping of channels.

To quantify these apparent differences in the response profiles, we performed two analyses (see **Methods**). First, we examined how quickly the S condition diverges from the W condition (**Figure 4A**), and how quickly the W condition diverges from the N condition (**Figure 4B**). The average

(across channels) classification performance revealed that channels in Cluster 1 reliably distinguish both S from W and W from N already in the earliest time bins. In contrast, channels in Cluster 2 reliably distinguish S from W and W from N starting at word positions 2-3 and onward, and channels in Cluster 3 do not reliably distinguish W from N in any continuous stretches of time and only distinguish S from W at the very last word position. And second, we examined how strongly the neural signal exhibits 'locking' to individual word/nonword onsets. This analysis revealed that—consistent with visual examination—channels in Cluster 3 show the strongest degree of stimulus locking, followed by channels in Cluster 2, with channels in Cluster 1 showing the weakest stimulus-related time-locking (**Figure 4C, Table S1).** This difference in the degree of stimulus locking was present across conditions, although the analysis additionally revealed a by-condition trend (**Figure 4C**) 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).

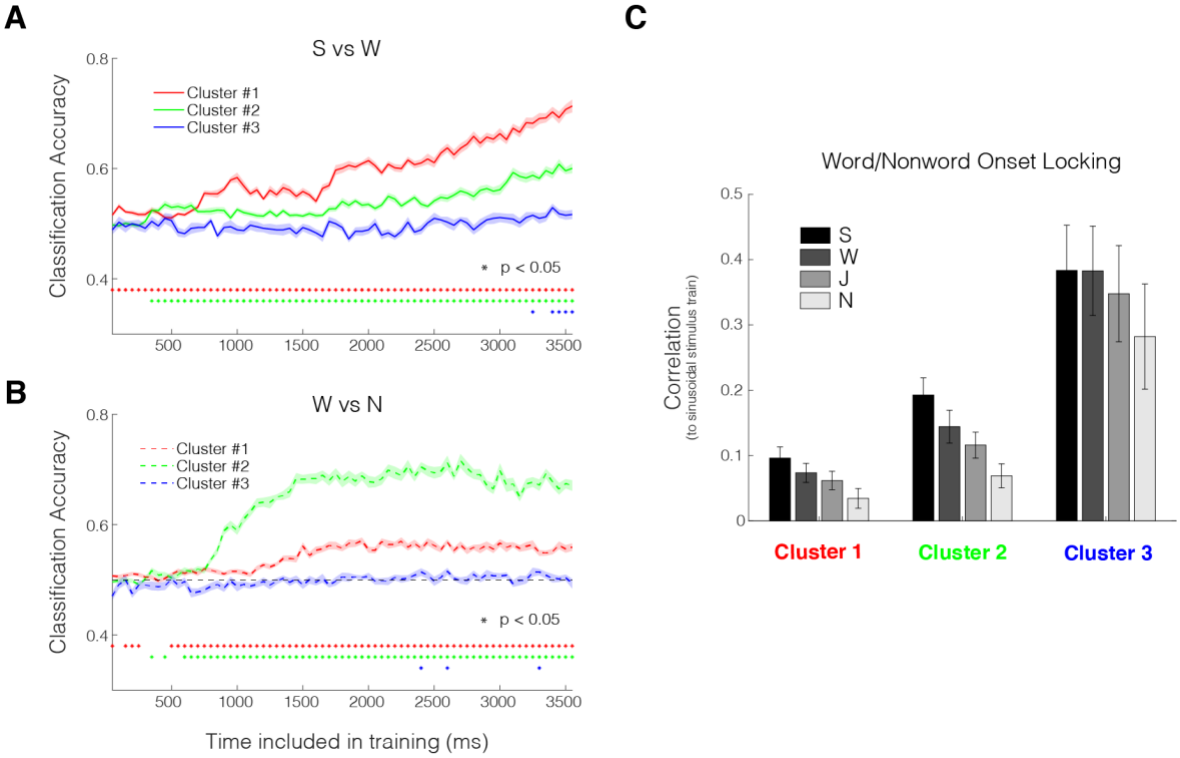

Figure 4 – Quantitative characterization of the three clusters in Dataset 1. A and B) Classifier performance by cluster as a function of the amount of timecourse included in training (**Methods**). Classifiers were trained to discriminate sentence and word list conditions (**A**) and word list and nonword list 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). The responses to sentences and word lists are most discriminable for channels in Cluster 1, whereas the responses to word lists and nonword lists are most discriminable for channels in Cluster 2. Responses to sentences and word lists as well as nonword lists cannot be discriminated (until the classifier has access to the entire timecourse in the case of discriminating sentences and word lists) for channels in Cluster 3 **C)** Correlation of fitted stimulus train with timecourse of channels by cluster and by condition (**Methods**). Channels in Cluster 3 are the most locked to word/nonword presentation while channels in Cluster 1 are the least locked to word/nonword presentation.

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 5A**). The clusters were also robust to the number of channels used: solutions derived from only a subset of the language-responsive channels (down to ~27%, ~32%, and ~69% of channels for clusters 1, 2, and 3, respectively) were significantly more similar to the clusters derived from all the channels than would be expected by chance (ps<0.05, **Methods**, **Figure 5B**). 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 channels (which is to be expected given that Cluster 3 has the fewest channels assigned to it, **Figure 3B**).

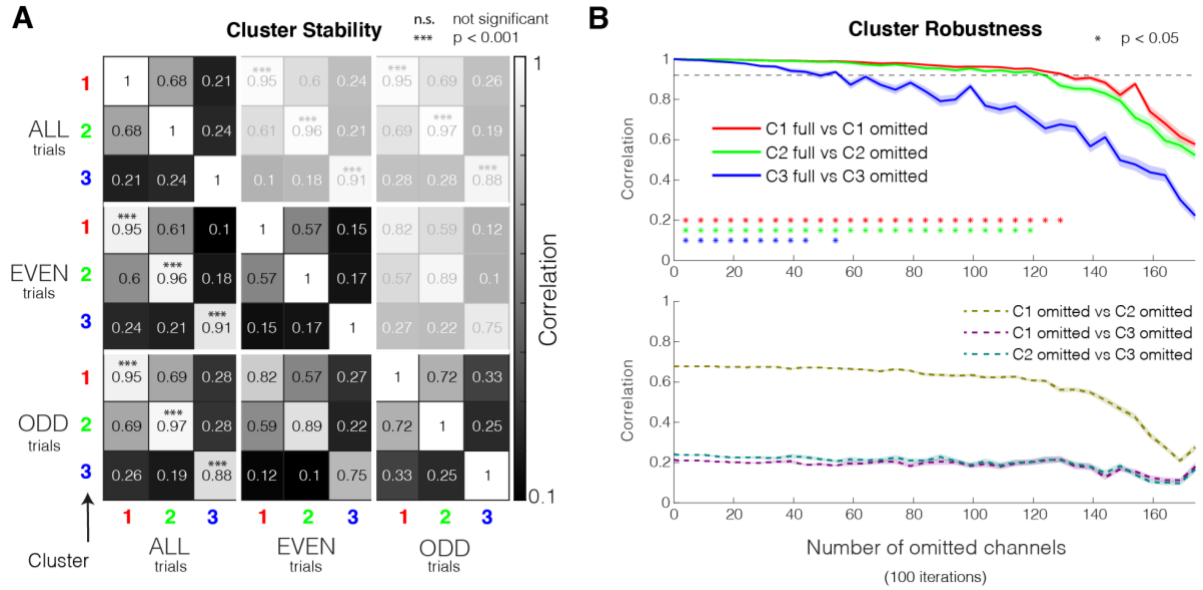

Figure 5 – Cluster evaluation within Dataset 1. A) Comparison of clusters from all trials (top three rows) versus only even (middle three rows) or odd (bottom three rows) trials. Clusters that emerge using only odd or even trials are strikingly similar to the clusters that emerge when all trials are used (ps<0.001; evaluated against clustering solutions when trials are shuffled; **Methods**). **B)** Robustness of clusters channel omission (random subsets of channels were removed in increments of 5). Top: Similarity of cluster centers when all channels were used versus when subsets of channels were removed. Stars reflect significant similarity with the full dataset (p<0.05; evaluated against clustering solution when trials are shuffled; **Methods**). Cluster 3 is driven the most by individual channels relative to Clusters 1 and 2. Bottom: Similarity between cluster centers when subsets of channels are removed. 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 channels (see Top).

Finally, we asked whether the same clusters would emerge in a second dataset with new participants and linguistic materials (Dataset 2; 16 participants; 362 language-responsive channels; **Figure 6A**). Because 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)), we first re-clustered the channels from Dataset 1 using only the

responses to the S and N conditions. The optimal number of clusters 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 S2**). In line with this similarity, ~80% of channels 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 channels were removed (**Figure S2G**) revealed that Cluster 2 was less robust than Clusters 1 and 3 to channel loss when only the S and N conditions were used (compare the green curve in **Figure 5B** vs. the green curve **Figure S2G**). 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 channels 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 S3A**), and the resulting clusters were reliably similar to Clusters 1 and 3 from Dataset 1 ($ps<0.001$ for both clusters, **Figure S3C**). However, to test whether Cluster 2 could emerge from Dataset 2, we also clustered Dataset 2 enforcing k=3. When the channels 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 6G,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 6G**), but the resemblance was not statistically reliable. This result may be, in part, attributable to the lower reliability of stimulus-evoked responses in Dataset 2 compared to Dataset 1 (**Figure 6B,C**), and the relatively sparse electrode coverage in Dataset 2 participants.

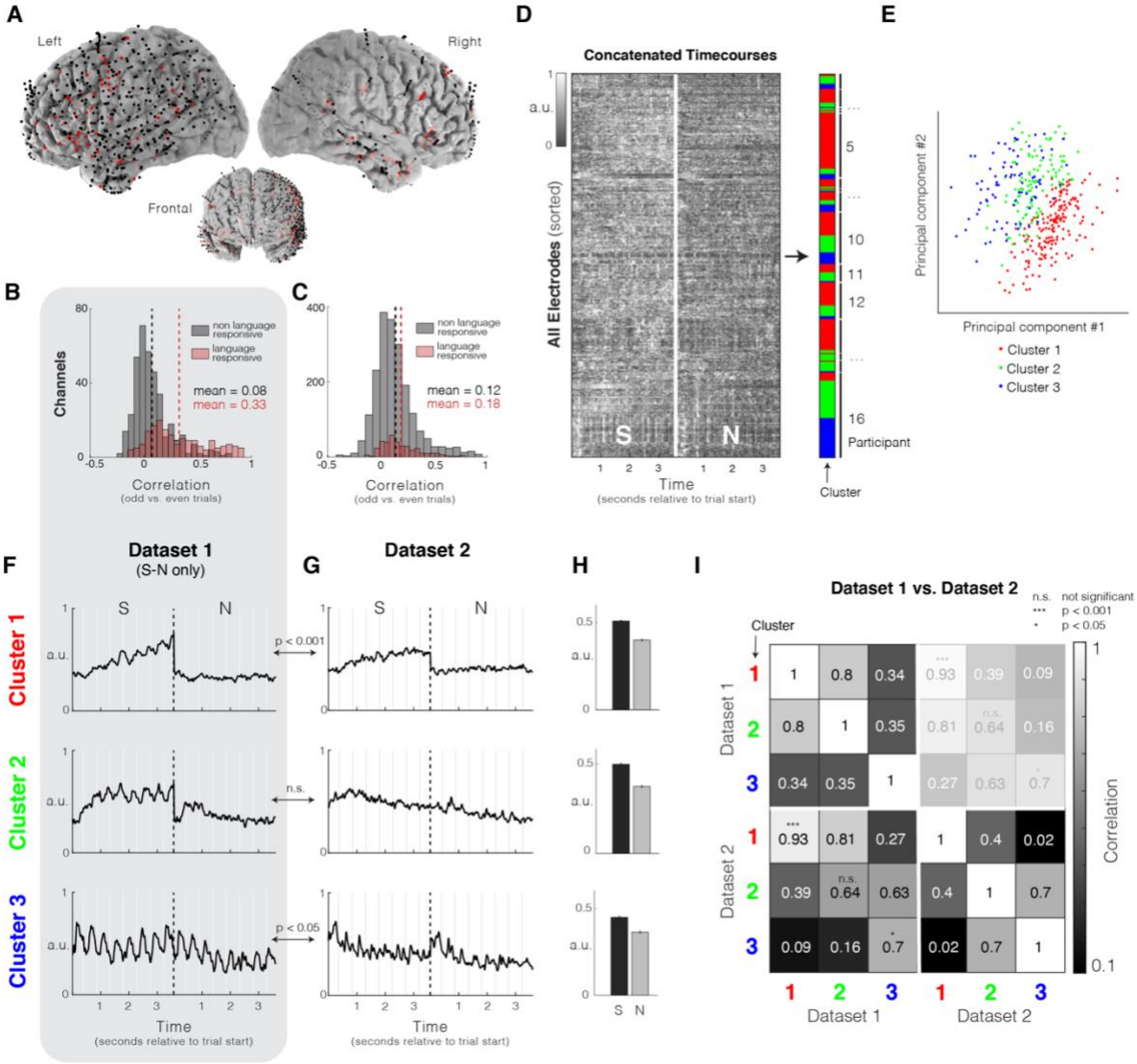

Figure 6 – Dataset 2 k-medoids clustering results (k=3). A) The locations of language-responsive (n=362, red; see Language-Responsive Channel Selection) and non-language-responsive (n=2,017, black) electrodes across the sixteen participants in Dataset 2 (both surface and depth electrode were implanted). Language-reponsive channels were found across the cortex, in both the left and right hemispheres (**Table 2**). **B and C)** Response stability across odd and even trials for language-responsive and non language-responsive channels. Language-responsive channels exhibit more reliable responses to linguistic stimuli than non language-responsive channels for both Dataset 1 (S+N conditions only, **B**) and Dataset 2 (**C**), however, the responses of language channels are less reliable in Dataset 2 than Dataset 1. **D)** Clustering mean channel 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. **E)** Channels 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. Again, after averaging across time, response profiles are not as distinct by cluster, underscoring the importance of temporal information in elucidating this grouping of channels. **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.

Thus, the three response profiles discovered in Dataset 1 were stable across trials and robust to data loss. Furthermore, two of the three response profiles were robustly detected in Dataset 1 using only two of the four conditions (sentences and nonword lists) and replicated in an independent larger dataset (Dataset 2) with new participants and new experimental materials.

2) 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 channels that belong to the 3 clusters in Dataset 1. We excluded from this analysis right-hemisphere (RH) channels because very few RH channels, 4 out of 177, passed the language selectivity criterion (S>N). We tested only the y (posterior-anterior) and z (inferior-superior) directions, but excluded the x (left-right) direction.

As can be seen from **Figure 7**, channels from both Cluster 1 and 2 were distributed across the temporal and frontal language regions. When examining all channels together or only the frontal channels, the MNI-coordinates of Clusters 1 and 2 did not significantly differ in either of the two tested directions (ps>0.05, LME, **Figure 7C, D**, **Methods, Table S2**). However, when examining only the channels located in the temporal lobe, channels from Cluster 1 were more inferior relative to Cluster 2 (p<0.001, LME, **Methods**, **Figure 7E, Table S2**), reflecting a large proportion of channels belonging to Cluster 1 on the ventral temporal surface.

Channels from Cluster 3 were more posterior than those in Clusters 1 and 2, manifesting as significant differences in the y-dimension (Cluster 3 vs. 1: p<0.001, LME, **Methods**, **Figure 7C, Table S2**; Cluster 3 vs. 2: p<0.01, LME, **Methods**, **Figure 7C, Table S2**). This trend was also evident when examining the temporal and frontal channels separately (**Figure 7D, E**), but the difference only reached significance for the temporal channels.

Dataset 2 was less suitable for investigating the anatomical distribution of the 3 clusters due to i) the lower data quality of Dataset 2 compared to Dataset 1 (compare **Figure 6B** vs. **6C**); ii) the non-reliable emergence of Cluster 2 in Dataset 2; and iii) the sparsity of coverage due to the prevalence of depth channels. Nevertheless, we attempted to apply a similar approach to the one described above to Dataset 2, focusing on Clusters 1 and 3, which emerged robustly in this dataset (**Figure S4**). The trends reported above (Cluster 3 channels located more posteriorly, and Cluster 1 channels more inferiorly) were weakly present but not reliable (**Figure S4**, **Table S3** and **Table S4**).

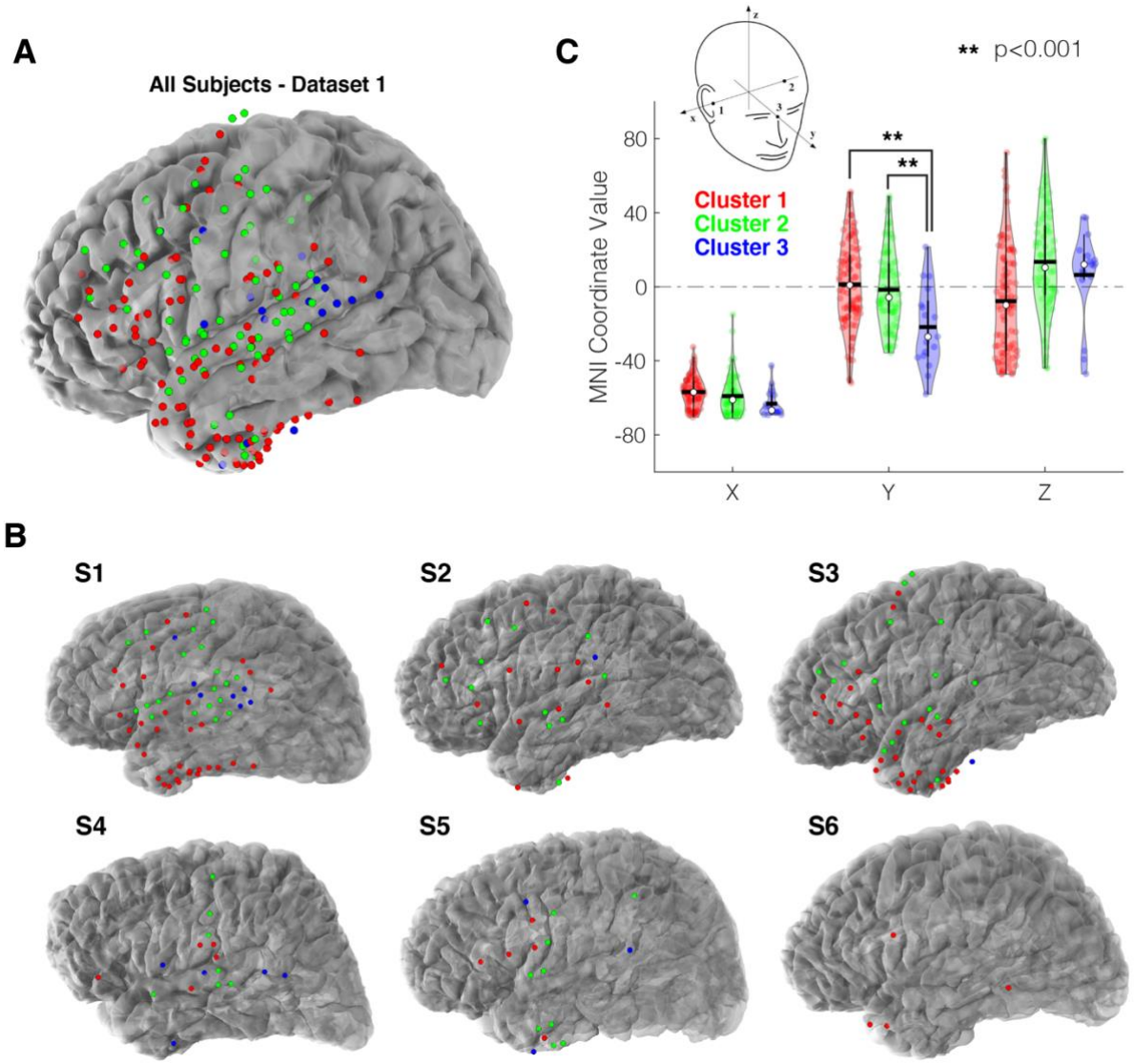

Figure 7 – Anatomical distribution of the clusters in Dataset 1. A) Anatomical distribution of language-responsive channels in Dataset 1 across all subjects in MNI space, colored by cluster. **B)** Anatomical distribution of language-responsive channels in subject-specific space. **C-E)** Violin plots 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 channels (**C**)**.** When channels 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**).

Discussion

What neural computations 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 distinct temporal profiles of response during the processing of sentences and linguistically degraded conditions like lists of words or nonwords. Further, we found that channels exhibiting these response profiles did not cluster spatially, 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

Our clustering approach revealed three distinct temporal profiles. We discovered these profiles in a dataset where participants processed 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). This "SWJN" design 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). In spite of some claims about regional dissociation between lexico-semantic and syntactic/combinatorial processing based on this and other paradigms(e.g., Dapretto and Bookheimer, 1999; Embick et al., 2000; Friederici et al., 2000; Kuperberg et al., 2000; see Fedorenko et al., 2020 for a recent review of this literature), studies that rely on robust individual-subject analyses have shown that these two aspects of language processing do not dissociate: all areas of the language network show a profile whereby the response is strongest to sentences, lower to lists of words and Jabberwocky sentences, and lowest to lists of nonwords (e.g., Fedorenko et al., 2010; Bedny et al., 2011; Shain et al., 2021; see Dick et al., 2001 for early arguments against the dissociation between lexical and syntactic processing). This ubiquitous sensitivity to both word meanings and syntactic structure building aligns with studies that have reported robust sensitivity to structure building across all parts of the language network (Blank et al., 2016; Fedorenko et al., 2020; Shain, Blank et al., 2020; Shain et al. 2022a) and supports views of language whereby sentence structure building is deeply intertwined with the processing of word meanings (e.g., Bybee, 1999, 2013; Goldberg, 2003; Jackendoff, 2007; Arnon and Snider, 2010; Jackendoff and Audring, 2020).

Importantly, however, although past data rule out hypotheses about the existence of neural mechanisms that are sensitive to linguistic structure and *not* to word meanings, many questions remain about how linguistic representations are built—the question the current study aims to contribute to. Using intracranial human recordings, Fedorenko et al. (2016) showed that a) the

overall pattern of response to sentences, word lists, Jabberwocky sentences, and nonwords lists mirrors that in fMRI studies and that b) no electrodes respond to structure in a lexical-content-independent manner (as would be evidenced by similar profiles for sentences and Jabberwocky sentences). They also 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) and they interpreted this component as indexing the process of constructing a sentence-level meaning. But these "build-up" channels constituted only a fraction of language-responsive channels. Here, we investigated the temporal profiles of language-responsive electrodes in a comprehensive manner. In particular, we identified language-responsive channels 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 (although we used a more liberal threshold to include as many potentially relevant channels as possible), and then used a dimensionality reduction approach to uncover distinct response profiles across all four conditions.

We uncovered three distinct profiles. One set of channels (~52% of the language-responsive channels) 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 channels (~38% of the language-responsive channels) 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 looked similar 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, and then decay). Finally, the remaining ~10% of the channels 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) whereas Cluster 2 in that dataset was only qualitatively similar to that in Dataset 1.

It is worth noting that the build-up channels reported in Fedorenko et al. (2016) likely represent a mix of channels from Clusters 1 and 2: the timecourse of response to the S condition in Fedorenko et al. is most similar to that in Cluster 1, but the fact that in that study, a reliable S>W>J>N profile was present suggests a contribution from Cluster 2 channels (given that Cluster 1 channels show similar-magnitude responses to conditions J and N rather than a J>N effect). As such, our analyses replicate the previous finding of the build-up effect (including in a new, larger dataset: Dataset 2), but refine it by identifying two functionally distinct build-up profiles.

Different channels may 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), which denote the amount of the preceding context that a given neural unit is sensitive to / 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 single word and a sentence (e.g., Lerner et al., 2011; Jacoby and Fedorenko, 2018; Blank and Fedorenko,

2020; Jain et al., 2020; Caucheteux and King, 2022). However, the lower end of the timescale (from single phonemes to syllables/morphemes to words) has not been much explored and some recent work has reported strong sensitivity in the language areas to sub-lexical regularities (Regev et al., 2021). One possibility therefore is that the language system comprises neural mechanisms that span the range from sub-lexical units (e.g., syllables) to single words to phrases and sentences. Our Clusters 3, 2, and 1, respectively, could correspond to these information integration windows.

To elaborate: two features of the clusters support the notion that channels across clusters differ in the scale over which they integrate information. First, a channel 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 channels in Cluster 3 (**Figure 4A**): these channels do not reliably discriminate between the S and W conditions, suggesting that these channels do not participate in combining words into larger units. The fact that they also do not reliably discriminate between the W and N conditions suggests that they are sensitive to spans of information below the word level. In contrast, a channel 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 channels in Clusters 2 and 1: these channels reliably discriminate between the S and W conditions.

Furthermore, channels 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 channels 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). Channels in Cluster 1 can already discriminate between the sentence and word-list conditions 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 very unlikely to start a phrase/sentence. Channels in Cluster 2 cannot discriminate between the sentence and word-list conditions until the second or third word, which may suggest that these channels only integrate information across a span of -2-3 words, and it takes until the second or third word to accumulate evidence that the words cannot be combined (after which point the gamma response declines; **Figures 3** and **4**). 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”. As a result, we conjecture that channels in Cluster 2 have the integration window of 2-3 words, and channels in Cluster 1 have the integration window of > 3 words (perhaps ~6 words; Shain et al., 2022).

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. A channel 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, a channel that processes information over spans of multiple words should exhibit a response that reflects a highly smoothed version of the stimulus train, with no momentary stimulus-related fluctuations. Indeed, we found that the three clusters differed 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 4C**).

Why do channels that are putatively sensitive to multi-word contexts (Cluster 1 channels) still show a stronger response to word-lists than nonword-lists (**Figure 3F**)? We can speculate that the ubiquity of the S>W(>)J>N pattern in the language-responsive channels is due to the strong functional interconnectivity of the language system (e.g., Blank et al., 2014; Braga et al., 2020), such that in spite of this apparent division of labor based on the size of the integration windows, all channels continuously share information, including feedback information from channels with longer TRWs to those with shorter ones. A natural future direction would be to study the extent to which channels within versus between clusters are functionally connected by computing temporal correlations between channels during naturalistic tasks.

It is important to note that the division of labor within the language system in terms of the scale of temporal integration is largely orthogonal to the putative dissociation between lexico-semantic and syntactic processing (discussed in the Introduction). In particular, channels that reliably discriminate between the Word-list and Nonword-list conditions (i.e., channels in Clusters 1 and 2; **Figure 4B**)—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 channels in Cluster 2 nor in Cluster 1 show responses to linguistic structure that is independent of word meaning processing, with much stronger responses to the Sentence condition compared to the Jabberwocky condition. These results support strong integration between the processing of word meanings and phrase-structure buildings as has been argued in past fMRI work (e.g., Fedorenko et al., 2020).

The distributed nature of language processing

There is a long history in the neuroscience of language 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 characterizes language processing, as evidenced by a clear separation 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) and b) putative higher-level areas that may carry out further processing on the meaning representations extracted from language (e.g., the Theory of Mind network; Saxe et al., 2006; see Paunov et al., 2019 and Shain et al., 2022b for evidence of a dissociation between these two systems, or see Ivanova, 2022 for dissociation between the

language and semantic demand regions). However, finding spatial sub-divisions within the language-selective fronto-temporal network has proven challenging (e.g., Fedorenko et al., 2010, 2020; Bautista & Wilson, 2016; Regev et al., 2021).

The current work demonstrates that there exists functional differentiation within the language network, but that 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 heterogeneity with respect to TRWs as discovered in an encoding approach with artificial neural network (ANN) language models).

Limitations and 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 (e.g., channels with short TRWs, but not with longer ones should be sensitive to features like word length and lexical frequency; in contrast, only channels with long TRWs should be sensitive to the difficulty of forming non-local dependencies). Second, ANN language models could be leveraged to better understand the constraints on language processing architecture, like whether successful language architectures require particular proportions of units with different TRWs or particular distributions of such units within and/or across model layers. In the current datasets, we found fewest channels belonging to Cluster 3 (shortest TRW), more channels belonging to Cluster 2 (longer TRW), and the majority of channels 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 channels tend to be implanted, so investigating these questions in ways where we can probe all composite units and have the freedom to alter the architecture in various ways is likely to yield insights that cannot be gained from human brains at least with the current tools available.

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 precise structure and content. We discovered three functionally distinct temporal profiles across language-responsive channels. We take these differences to reflect across-channel differences in the size of the temporal receptive window, with some channels paying attention to only local structure and others integrating information across words. However, in order for the language system to extract meaning from the signal, all three types of channels have to also represent the precise linguistic content. Because of the relatively sparse coverage of the intracranial recording approach, it may be easier to discover features that characterize content-independent features of the response profile (like the TRWs). However, 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 (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 specific linguistic meanings.

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 U S A 104:9087–9092.

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 U S A 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.

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, King JR (2022) Brains and algorithms partially converge in natural language processing. Commun Biol 2022 51 5:1–10.

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 U S A 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.e9.
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 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 U S A 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 U S A 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.
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. Nat 2016 5327600 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. 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.
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.
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 Allen P, ed. 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.
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. 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 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.
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.
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.
Pallier C, Devauchelle AD, Dehaene S (2011) Cortical representation of the constituent structure of sentences. Proc Natl Acad Sci U S A 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 U S A 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.
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

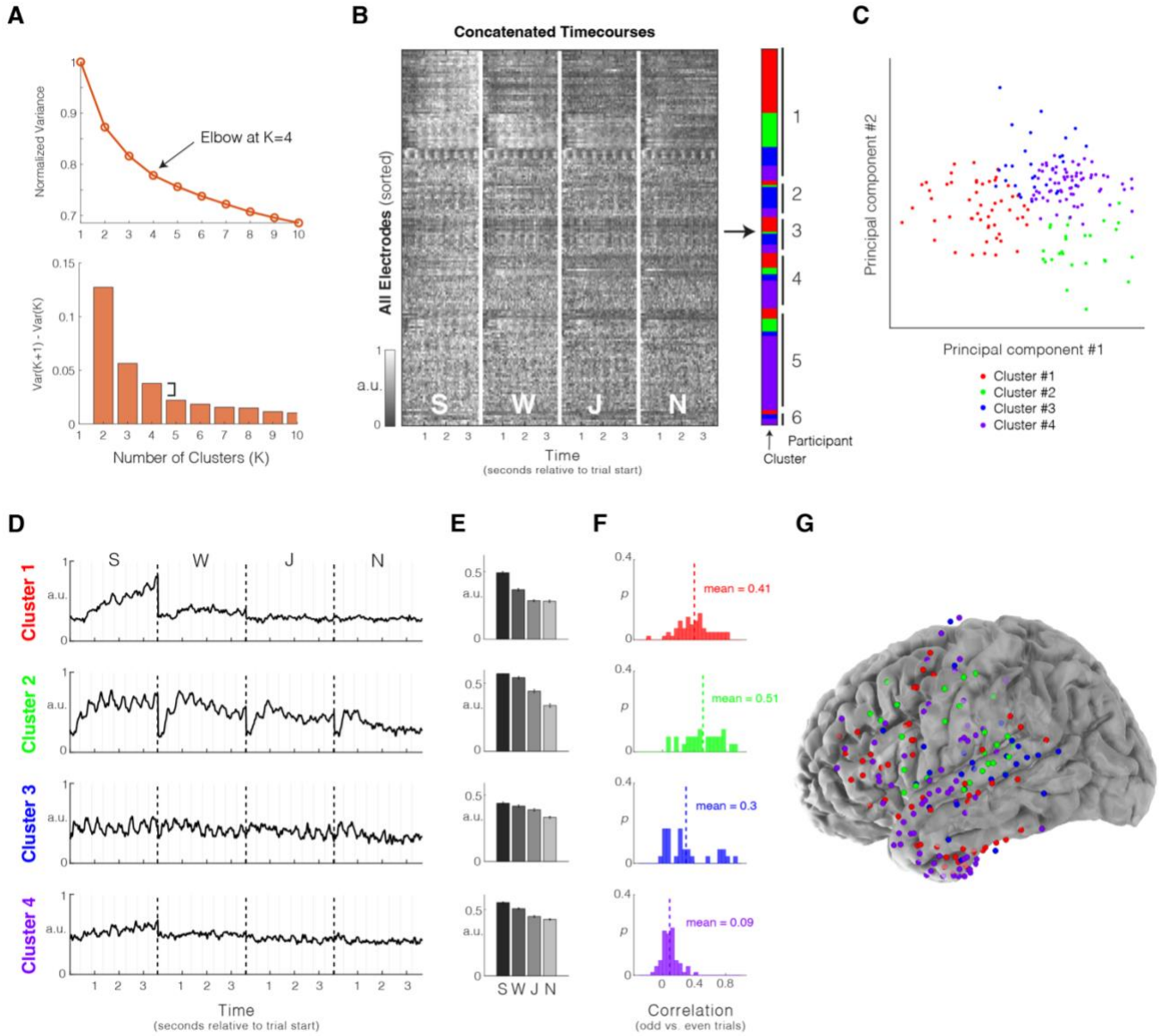

Figure S1 – Dataset 1 k-means clustering results. A) Search for optimal k using the "elbow method". Top: variance normalized by the variance when k=1 as a function of k. Bottom: change in normalized variance as a function of k. After k=4 the change in variance becomes more moderate, suggesting that 4 clusters appropriately describe this dataset. **B)** Clustering mean channel responses (S+W+J+N) using k-means (k=4) with squared-Euclidean distance. Shading of the data matrix reflects normalized high-gamma power. **C)** Channel 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. **F)** Response stability across odd and even trials by cluster. Channels 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).

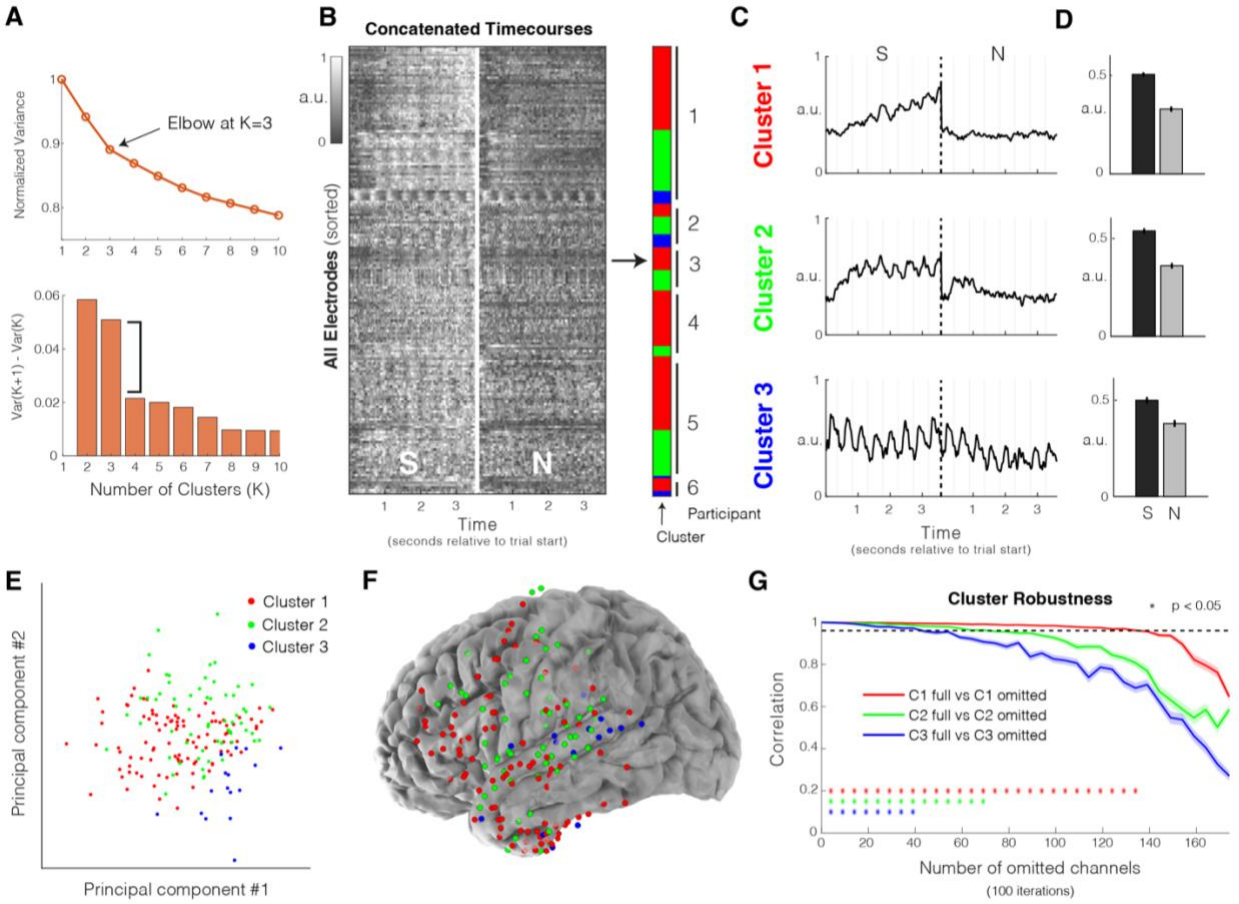

Figure S2 – Dataset 1 k-medoids clustering results with only S-N conditions. A) Search for optimal k using the "elbow method". Top: variance normalized by the variance when k=1 as a function of k. Bottom: change in normalized variance as a function of k. After k=3 the change in variance becomes more 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 channel 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. **C)** Average timecourse by cluster. Clusters 1-3 show striking similarity to the clusters reported in **Figure 3**. **D)** Mean condition responses by cluster. **E)** Channel 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 channel omission (random subsets of channels 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**). Relative to when all conditions were used, Cluster 2 is less robust to channel 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 channels from Cluster 2 from channels from Clusters 1 and 3.

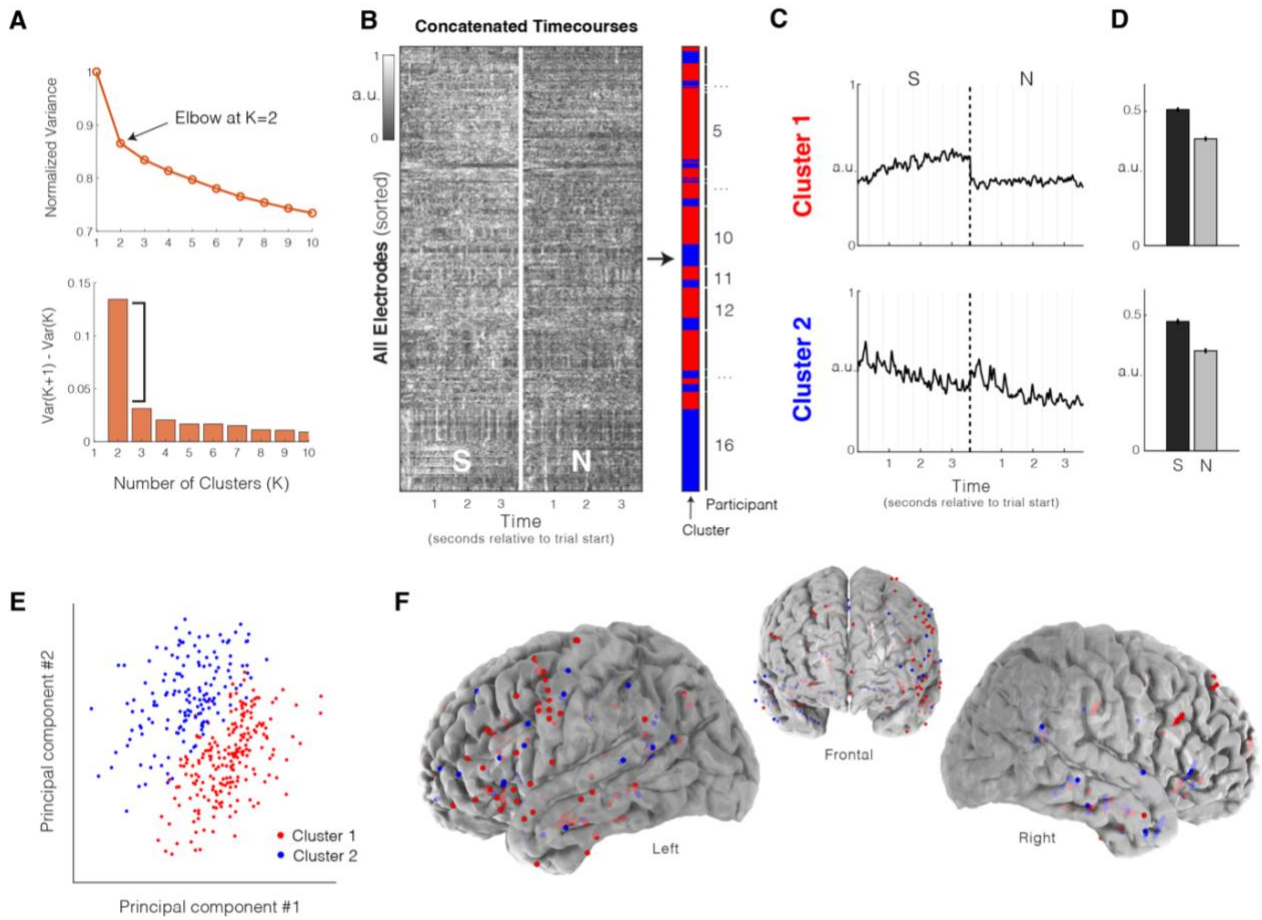

Figure S3 – Dataset 2 k-medoids clustering results with k=2. A) Search for optimal k using the "elbow method". Top: variance normalized by the variance when k=1 as a function of k. Bottom: change in normalized variance as a function of k. After k=2 the change in variance becomes more moderate, suggesting that 2 clusters appropriately describe Dataset 2 (for a direct comparison of Dataset 1 and Dataset 2 when k=3, see **Figure 6**). **B)** Clustering mean channel responses (S+N) using k-medoids (k=2) with a correlation-based distance. Shading of the data matrix reflects normalized high-gamma power. **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. **E)** Channel responses visualized on their first two principal components, colored by cluster. **F)** Anatomical distribution of clusters across all participants (n=16).

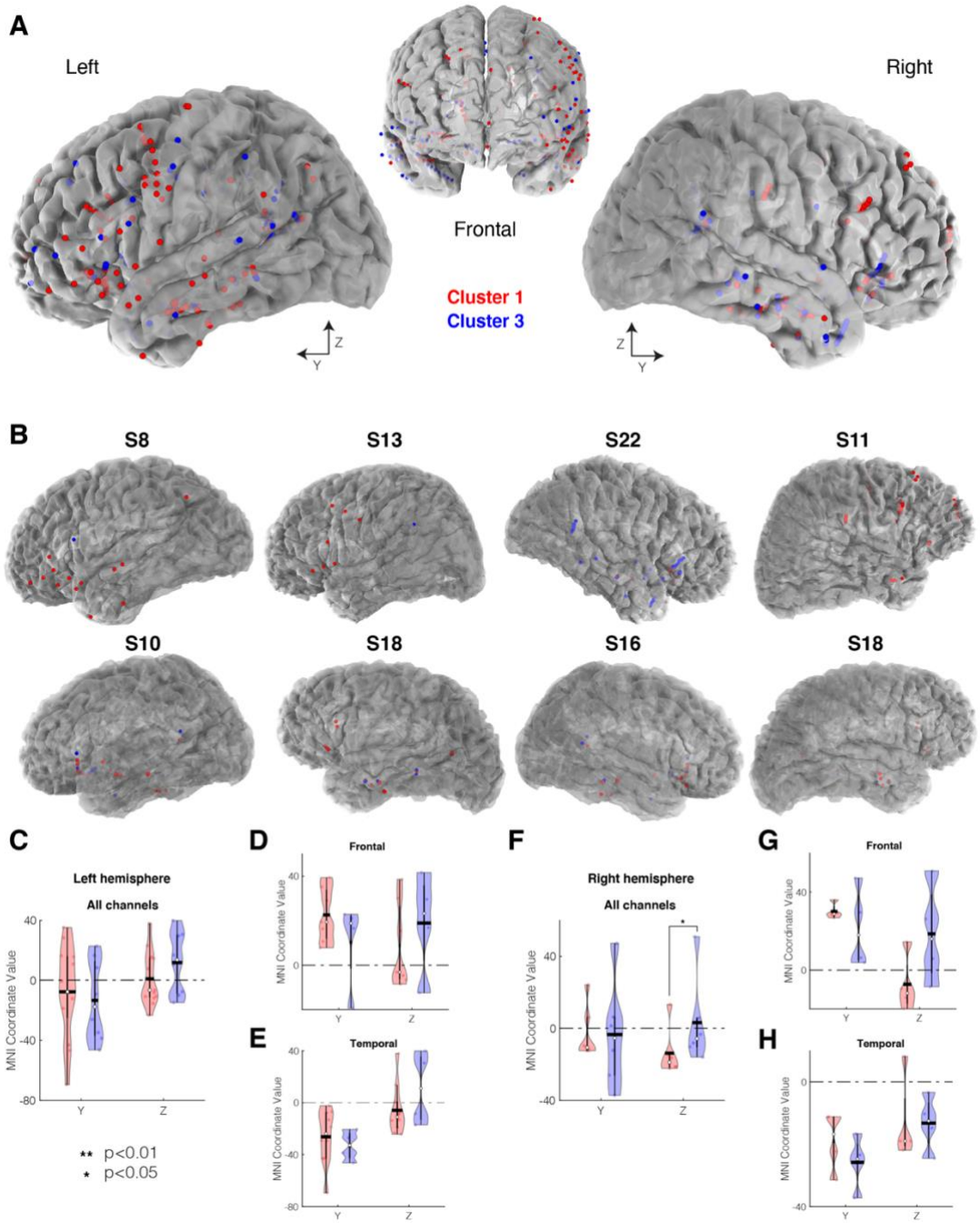

Figure S4 – Anatomical distribution of the clusters in Dataset 2. Anatomical distribution of language-responsive channels in Dataset 2 across all subjects in MNI space, colored by cluster. Only Clusters 1 and 3 (those from Dataset 1 that replicate to Dataset 2) are shown. **B)** Anatomical distribution of language-responsive channels 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.

Name	Estimate	SE	tStat	DF	pValue
(Intercept)	0.102	0.022	4.651	696	4.0E-06
k3_2	0.077	0.035	2.180	696	3.0E-02
k3_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_k3-2	NaN	NaN	2.14	1	0.03
coefTest_N-W	NaN	NaN	1.45	1	0.15
coefTest_J-W	NaN	NaN	0.50	1	0.62
coefTest_N-J	NaN	NaN	0.85	1	0.40
k3_2:condition_W	-0.03	0.04	-0.66	696	0.51
k3_3:condition_W	0.02	0.06	0.35	696	0.73
k3_2:condition_J	-0.04	0.04	-1.06	696	0.29
k3_3:condition_J	0.01	0.06	0.10	696	0.92
k3_2:condition_N	-0.06	0.04	-1.60	696	0.11
k3_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 on the variables cluster and condition (**Methods**). Model formula: *Locking ~ 1 + cluster*condition + (cluster|participant) + (condition|participant)*). 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.

Dataset	Coordinate	Name	Estimate	SE	tStat	DF	pValue
Dataset 1	x	Cluster 1	-57.06	1.21	-47.04	171	9.2E-100
Dataset 1	x	Cluster 2	-2.12	1.53	-1.38	171	0.16845
Dataset 1	x	Cluster 3	-4.56	2.40	-1.90	171	0.05907
Dataset 1	x	coefTest_Cluster3-2	NaN	NaN	-1.02	1	0.31140
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 coordinates of the 3 clusters, Dataset 1, Left hemisphere.

Dataset	Coordinate	Name	Estimate	SE	tStat	DF	pValue
Dataset 2	x	Cluster 1	-44.89	3.25	-13.81	196	9.5E-31
Dataset 2	x	Cluster 2	-0.57	5.15	-0.11	196	0.9125
Dataset 2	x	Cluster 3	3.32	3.62	0.92	196	0.3599
Dataset 2	x	coefTest_Cluster3-2	NaN	NaN	0.67	1	0.5031
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.

Dataset	Coordinate	Name	Estimate	SE	tStat	DF	pValue
Dataset 2	x	Cluster 1	35.0	3.1	11.3	160	3.5E-22
Dataset 2	x	Cluster 2	9.9	5.4	1.9	160	0.066
Dataset 2	x	Cluster 3	1.7	6.1	0.3	160	0.776
Dataset 2	x	coefTest_Cluster3-2	NaN	NaN	1.0	1	0.331
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.