nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01944-2

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

Received: 16 March 2023

Accepted: 3 July 2024

Published online: 26 August 2024

Check for updates

Tamar I. Regev[1,2,9] ✉, Colton Casto[1,2,3,4,9] ✉, Eghbal A. Hosseini[1,2], Markus Adamek[5,6], Anthony L. Ritaccio[7], Jon T. Willie[5,6], Peter Brunner[5,6,8] & Evelina Fedorenko[1,2,3] ✉

Despite long knowing what brain areas support language comprehension, our knowledge of the neural computations that these frontal and temporal regions implement remains limited. One important unresolved question concerns functional differences among the neural populations that comprise the language network. Here we leveraged the high spatiotemporal resolution of human intracranial recordings (n = 22) to examine responses to sentences and linguistically degraded conditions. We discovered three response profiles that differ in their temporal dynamics. These profiles appear to reflect different temporal receptive windows, with average windows of about 1, 4 and 6 words, respectively. Neural populations exhibiting these profiles are interleaved across the language network, which suggests that all language regions have direct access to distinct, multiscale representations of linguistic input—a property that may be critical for the efficiency and robustness of language processing.

Language processing engages a network of brain regions that reside in the temporal and frontal lobes and are typically left lateralized[1,2]. These brain regions respond strongly to linguistic stimuli across presentation modalities[1,3,4], tasks[1,5] and languages[6]. This language-responsive network is highly selective for language, showing little or no response to diverse non-linguistic inputs and tasks[7–11] (see ref. 12 for a review). However, the precise computations and neuronal dynamics that underlie language comprehension remain debated.

On the basis of neuroimaging and aphasia evidence, some have argued for dissociations among different aspects of language, including phonological/word-form processing[13–15], the processing of word meanings[16,17] and syntactic/combinatorial processing[18–21]. However, other studies have reported distributed sensitivity to these aspects of language across the language network[1,22–25]. Some of the challenges in discovering robust functional differences within the language network may have to do with the limitations of functional magnetic resonance imaging (fMRI)—the dominant methodology available for studying language processing. Each fMRI 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, the relatively slow temporal resolution of fMRI (typically, ~2 s) may obscure the dynamics of linguistic computations.

In recent years, invasive recordings of human neural activity[26], including electrocorticography (ECoG) and stereo electroencephalography (sEEG), have become increasingly available to language

[1]Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Program in Speech and Hearing Bioscience and Technology (SHBT), Harvard University, Boston, MA, USA. [4]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Allston, MA, USA. [5]National Center for Adaptive Neurotechnologies, Albany, NY, USA. [6]Department of Neurosurgery, Washington University School of Medicine, St Louis, MO, USA. [7]Department of Neurology, Mayo Clinic, Jacksonville, FL, USA. [8]Department of Neurology, Albany Medical College, Albany, NY, USA. [9]These authors contributed equally: Tamar I. Regev, Colton Casto. ✉e-mail: tamarr@mit.edu; ccasto@mit.edu; evelina9@mit.edu

https://doi.org/10.1038/s41562-024-01944-2

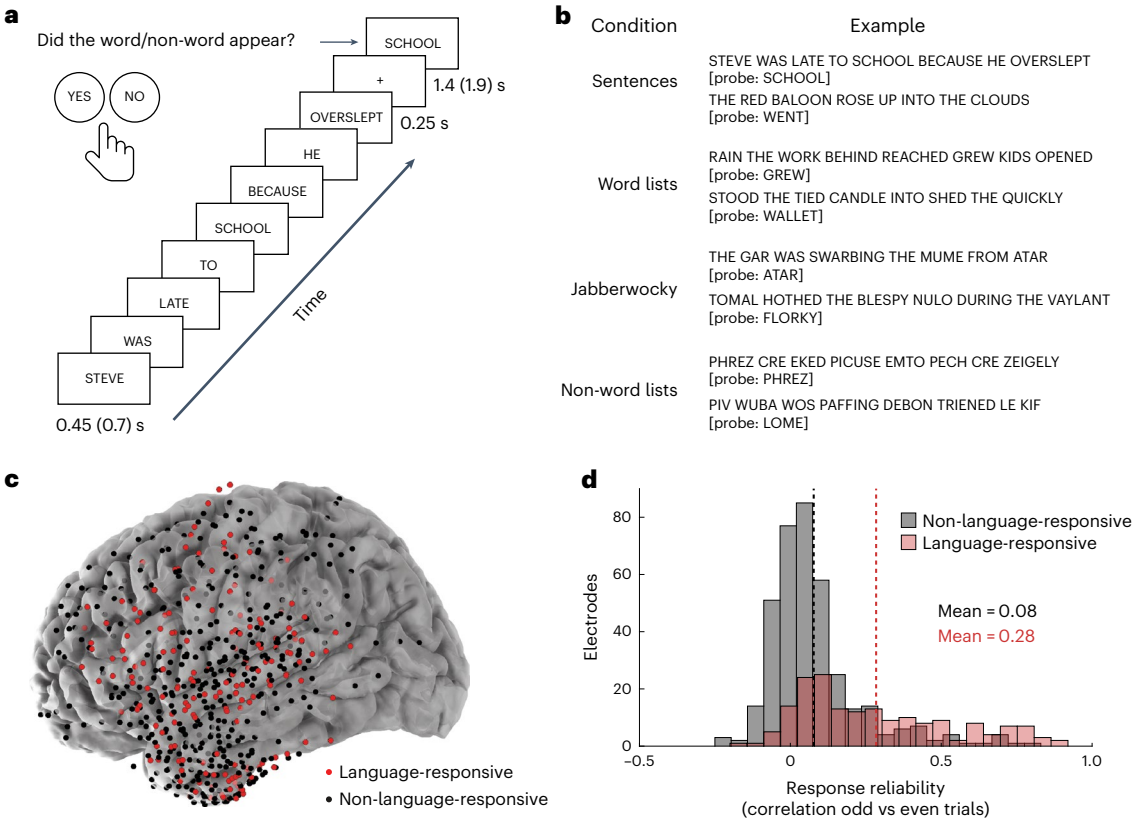


Fig. 1 | Experimental procedure and the distribution of the implanted electrodes for Dataset 1. a, A sample trial from the Sentence condition. **b**, For each of the four experimental conditions, items are presented with word/non-word probes that either appeared in the trial or not. Adapted from ref. 27. **c**, The locations of language-responsive (n = 177, red; Methods) and non-language-responsive (n = 373, black) electrodes across the 6 participants in Dataset 1. Electrodes were implanted almost exclusively in the left hemisphere for Dataset 1 and concentrated in the temporal and frontal lobes. **d**, Response reliability across odd and even trials (based on a correlation of average condition-level responses) for language-responsive and non-language-responsive electrodes. Language-responsive electrodes exhibit more reliable responses to linguistic stimuli than non-language-responsive electrodes. Dashed vertical lines represent the mean of the corresponding distribution.

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 and temporal resolution, allowing the tracking of neural dynamics across both space and time. Several previous studies have probed intracranial neural responses during language comprehension[27–31]. For example, ref. 27 reported sensitivity in language-responsive electrodes to both word meanings and combinatorial processing, in line with fMRI findings (for example, ref. 1). They also reported a temporal profile where neural activity gradually increases (builds up) across the sentence (replicated in refs. 28–30), which they interpreted as reflecting the construction of a sentence meaning. However, considerable disagreement exists in the field regarding the number of distinct profiles that characterize cortical language responses, how they functionally differ and what computations they collectively support in the service of language comprehension and production.

Here we report a detailed investigation of neural responses during language processing. To isolate the language network from nearby lower-level perceptual areas and domain-general cognitive areas, we focus on electrodes that show a characteristic functional signature of the language areas: a stronger response to sentences than to sequences of non-words (as in ref. 27). To foreshadow our findings, we report three response profiles that differ in their temporal dynamics and overall magnitude of response to linguistically degraded conditions. Using a toy model with a single parameter—the timescale of information integration—we argue that these profiles reflect distinct temporal receptive window sizes in the language network[32–34].

Results

We used intracranial recordings from patients with intractable epilepsy to investigate neural responses during language comprehension. Participants in Dataset 1 were presented with four types of linguistic stimuli that have been traditionally used to tease apart neural responses to word meanings and syntactic structure[1,2,27,30,35,36]: sentences (S), lists of unconnected words (W), Jabberwocky sentences (where content words were replaced with non-words, J) and lists of unconnected non-words (N) (Fig. 1a,b and Methods, all stimuli are available on OSF[37]). In each trial, 8 words or non-words were presented on a screen serially and participants were asked to silently read them. To maintain alertness, after each trial, participants judged whether a probe word/non-word had appeared in that trial. See Methods for further details of stimulus presentation and behavioural response data. In Dataset 2, just two of these conditions were used: sentences and lists of non-words.

We asked three research questions: (1) Does the language network contain reliably distinct response profiles? If so, (2) What do these profiles reflect? And finally, (3) Do electrodes exhibiting different response profiles tend to be located in particular regions of the language network? We used Dataset 1 (n = 6) for initial evaluation of these questions because this dataset contained a richer set of experimental conditions. We then used Dataset 2 (n = 16) as an attempt to replicate the findings despite the more compact experimental paradigm.

Language electrodes exhibit distinct response profiles

We clustered the high-gamma neural response patterns of language-responsive electrodes from Dataset 1 (6 participants, same as

https://doi.org/10.1038/s41562-024-01944-2

Table 1 | Details for Dataset 1

Participants	Age	Sex	Site	ECoG or sEEG	Language-responsive electrodes (S>N)	Total clean electrodes	Total electrodes	Native sampling freq. (Hz)	Elec. per amp.	Runs	Pres. rate (per word)	Trials per cond.
Participant 1	29	F	AMC	ECoG	62 (0 RH)	108 (0 RH)	120 (0 RH)	1,200	16	10	450 ms	80
Participant 2	25	F	AMC	ECoG	17 (0 RH)	115 (0 RH)	128 (0 RH)	1,200	16	10	700 ms	60
Participant 3	18	F	AMC	ECoG	17 (0 RH)	92 (0 RH)	98 (0 RH)	1,200	16	10	700 ms	60
Participant 4	28	M	AMC	ECoG	26 (0 RH)	106 (0 RH)	134 (0 RH)	1,200	64	10	700 ms	60
Participant 5	25	F	AMC	ECoG	48 (0 RH)	93 (0 RH)	98 (0 RH)	1,200	64	10	450 ms	80
Participant 6	20	F	AMC	ECoG	7 (3 RH)	36 (20 RH)	36 (20 RH)	1,200	64	10	450 ms	80

All data were collected at the Albany Medical Center (AMC). Here and in Table 2, 'Total electrodes' excludes reference electrodes, ground electrodes, microphone electrodes, trigger electrodes, skull EEG electrodes and EKG electrodes; and 'Total clean electrodes' excludes electrodes with significant line noise, significant interictal discharges, or large visual artefacts identified through manual inspection. 'Elec. per amp.' is the number of electrodes per amplifier. 'Pres. rate (per word)' is the duration of presentation for each single word or non-word.

those used in ref. 27, 177 language-responsive electrodes; Fig. 1c, Methods and Table 1) to sentences (S), word lists (W), Jabberwocky sentences (J) and non-word lists (N) (Fig. 1a,b). We focused on differences across experimental conditions and therefore, clustering was performed on the average condition timecourses, which were concatenated across the four conditions to create a single timecourse per electrode (Fig. 2b and Methods). The *k*-medoids clustering algorithm, combined with the 'elbow' method (Methods), suggested that three clusters ($k = 3$) optimally explain the data (Fig. 2a; similar results emerged with a *k*-means clustering algorithm, see OSF[37]). Although we combined the electrodes from all 6 participants for clustering, electrodes that belong to each of the three clusters were evident in every participant individually (Fig. 2b and Extended Data Fig. 1).

Additional analyses further suggested that these three response types capture a substantial amount of the functional heterogeneity in our dataset. First, we repeated the clustering analysis while omitting electrodes below a parametrically varying reliability threshold and found that the elbow at $k = 3$ became more pronounced (Fig. 2a inset). Reliability was defined as the correlation between average, condition-level responses to odd versus even trials (Fig. 1d). Second, when clustering was performed using a larger value of *k* (for example, $k = 10$), the profiles of many of the additional clusters resembled the profiles that we discovered when clustering using $k = 3$ (Extended Data Fig. 2). And third, responses within a given cluster, especially the more reliable responses, appeared visually similar to the prototypical cluster response profiles, with only a couple of highly reliable responses exhibiting a distinct profile (Extended Data Fig. 3).

The average timecourses for the three clusters are shown in Fig. 2e (see Fig. 2d for most representative electrodes from each cluster ('medoids') chosen by the *k*-medoids algorithm). Cluster 1 ($n = 92$ electrodes (52% of all language electrodes); number of electrodes present in individual participants: 5–34, Extended Data Fig. 1) was characterized by a relatively slow increase (buildup) of neural activity across the 8 words in the S condition (a pattern similar to the one reported in refs. 27–30; but see Discussion), and much lower activity for the W, J and N conditions, with no qualitative difference between the J and N conditions (Fig. 2f). Cluster 2 ($n = 67$ electrodes (38% of all language electrodes); number of electrodes present in individual participants: 1–21, Extended Data Fig. 1) displayed a quicker buildup of neural activity in the S condition that plateaued approximately 3 words into the sentence, a quick build-up of activity in the W condition that began to decay after the third word, and a similar response to the J and N conditions as to the W condition with an overall lower magnitude. Cluster 2 also exhibited 'locking' of the neural activity to the onsets of individual words in the S condition. Finally, Cluster 3 ($n = 18$ electrodes (10% of all language electrodes); number of electrodes present in individual participants: 1–7, Extended Data Fig. 1) showed no buildup of activity and was instead characterized by a high degree of locking to the onset of each word or non-word in all conditions. In addition, the response magnitudes of Cluster 3 were more similar across conditions compared with the other two clusters, although the S > W > J > N pattern was still visually apparent (Fig. 2f).

We then evaluated the stability of these clusters across trials and their robustness to data loss. We found that clusters derived from half of the data (either odd- or even-numbered trials) were more similar to the clusters derived from the full dataset or from the other half of the data than would be expected by chance ($P < 0.001$, one-sided permutation test, $n = 1{,}000$ permutations; Methods and Fig. 3a). The clusters were also robust to the number of electrodes used: clustering solutions derived from only a subset of the language-responsive electrodes (down to ~27%, ~32% and ~69% of electrodes for Clusters 1, 2 and 3, respectively) were more similar to the clusters derived from all the electrodes than would be expected by chance (using a *P* threshold of 0.05, evaluated with a one-sided permutation test, $n = 1{,}000$ permutations; Methods and Fig. 3b).

To further quantify the apparent differences among the three response profiles, we performed two additional analyses. First, we examined how strongly the neural signal exhibited 'locking' to individual word/non-word onsets by correlating the observed responses with a fitted sinusoidal function (Methods). This analysis revealed that, consistent with visual examination, electrodes in Cluster 3 showed the strongest degree of stimulus locking, followed by electrodes in Cluster 2, with electrodes in Cluster 1 showing the weakest stimulus-related locking (significant overall effect of cluster on locking, analysis of variance (ANOVA) for linear mixed-effects models $F(2,9.13) = 5.4$, $P = 0.028$; Methods and Fig. 3c, see Supplementary Table 1a,b for a complete description of the statistical details and results). And second, we tested how quickly and strongly the S, W, J and N conditions diverged from one another in each of the profiles. We did this using a binary logistic classifier, trained for each cluster separately, using incrementally more of the timecourse for discrimination (Fig. 3d–f and Methods). Significance was evaluated as a one-sided cluster statistic against a null distribution from permuted labels (refs. 38,39, $n = 1{,}000$ permutations; Methods). The classification performance (averaged across 10 folds of the cross-validated classifier) revealed that neural populations in Cluster 1 reliably distinguished S from W earlier and more strongly than the neural populations in Clusters 2 and 3. In contrast, neural populations in Cluster 2 reliably distinguished W from N and J from N earlier and more strongly than neural populations in Clusters 1 and 3.

Although the *k*-medoids clustering algorithm assigns each electrode to one of *k* discrete clusters, we wanted to additionally evaluate the degree to which single electrode profiles fell between the prototypical cluster response profiles. To do this, we computed the partial correlation of every electrode's response profile with that of each of the cluster medoids while controlling for the other two medoids (Extended Data Fig. 4 and Methods). As shown in Extended Data Fig. 4b, many of the electrodes exhibited response profiles that were consistent with only 'one' of the prototypical responses. However, a few electrodes,

https://doi.org/10.1038/s41562-024-01944-2

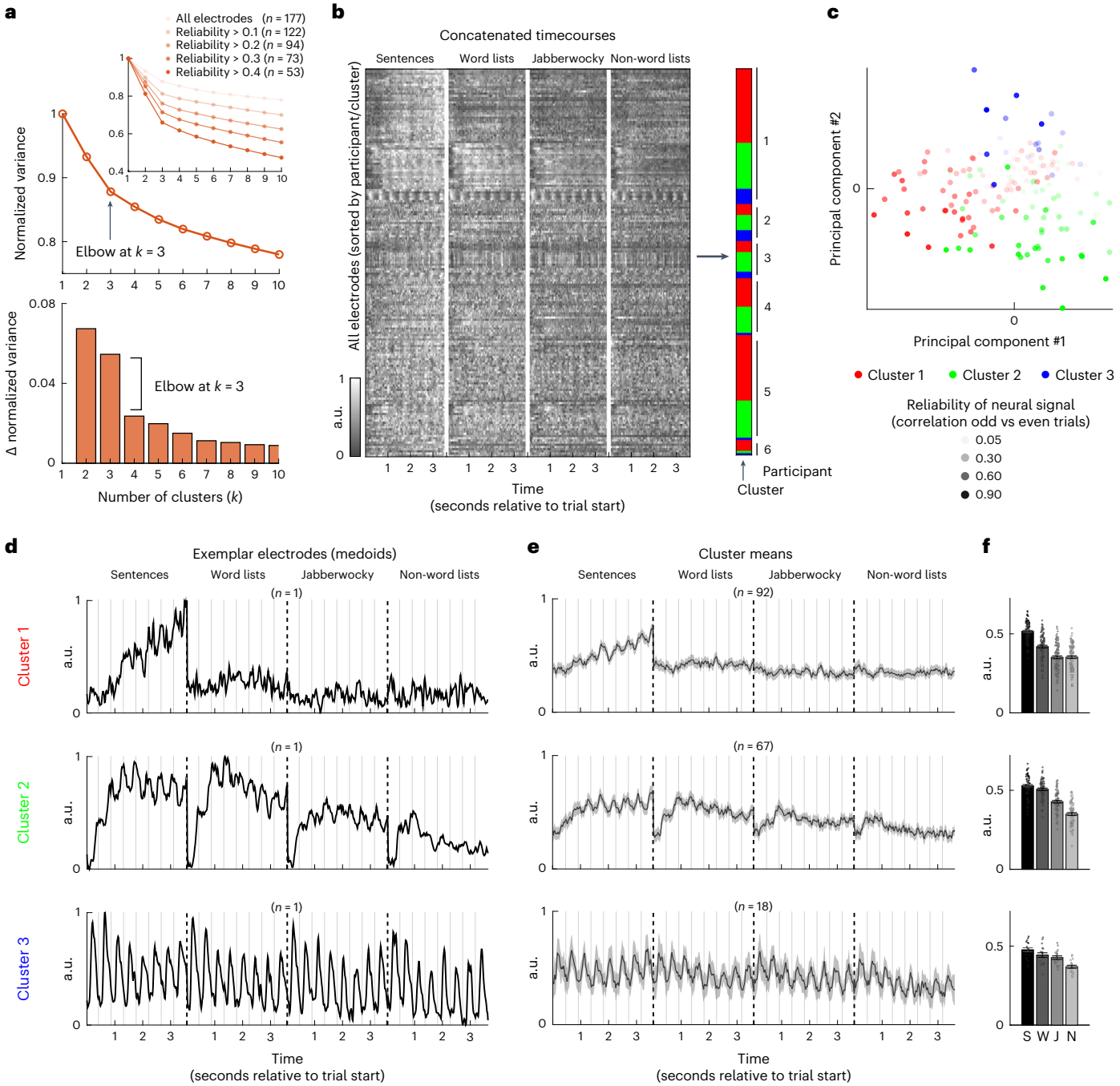


Fig. 2 | Dataset 1, *k*-medoids clustering with $k = 3$. a, Search for optimal k using the 'elbow method'. Top: variance (sum of the distances of all electrodes to their assigned cluster centre) normalized by the variance when $k = 1$, as a function of k (normalized variance, NV). Inset: clustering was performed while omitting electrodes below a parametrically sampled reliability threshold. Orange shading represents the reliability threshold for omitting electrodes. The elbow (point of transition between a steeper to a more moderate slope) gets more pronounced when eliminating lower-reliability electrodes, which suggests that $k = 3$ best describes these data. Bottom: change in NV as a function of k (NV($k + 1$) − NV(k)). After $k = 3$, there was a large drop in the change in variance. **b**, Clustering mean electrode responses (concatenated across the four experimental conditions: sentences (S), word lists (W), Jabberwocky (J), non-word lists (N)) using k-medoids ($k = 3$) with a correlation-based distance (Methods). Shading of the data matrix reflects normalized high-gamma power (70–150 Hz). Electrodes are sorted vertically by participant and their assignment to clusters (right colour bar). All three clusters are present in each of the 6 participants. **c**, Electrode responses visualized on their first two principal components, coloured by cluster and shaded by the reliability of the neural signal as estimated by correlating responses to odd and even trials (Fig. 1d). **d**, Timecourses of best representative electrodes ('medoids') selected by the algorithm from each of the three clusters. The timecourses reflect normalized high-gamma (70–150 Hz) power averaged over all trials of a given condition. a.u. stands for arbitrary units; the signals were z-scored and normalized to have minimum value of 0 and maximum value of 1. **e**, Timecourses averaged across all electrodes in each cluster ($n = 92$, 67 and 18 for Clusters 1, 2 and 3, respectively). Shaded areas around the signal reflect the 99% confidence interval over electrodes. **f**, Mean condition responses by cluster. Error bars reflect s.e.m. over electrodes ($n = 92$, 67 and 18 for Clusters 1, 2 and 3, respectively, as in **e**). Data points reflect individual electrodes. After averaging across time, response profiles are not as distinct by cluster (especially for Clusters 2 and 3), which underscores the importance of temporal information in elucidating this grouping of electrodes.

https://doi.org/10.1038/s41562-024-01944-2

Table 2 | Details for Dataset 2

Participant	Age	Sex	Site	ECoG or sEEG	Language-responsive electrodes (S>N)	Total clean electrodes	Total electrodes	Native sampling freq. (Hz)	Elec. per amp.	Runs	Pres. rate (per word)	Trials per cond.
Participant 7	51	M	AMC	ECoG	14 (7 RH)	116 (25 RH)	126 (26 RH)	1,200	64	3	750 ms	48
Participant 8	30	F	AMC	both	18 (0 RH)	76 (1 RH)	92 (3 RH)	1,200	64	3	750 ms	72
Participant 9	31	M	AMC	sEEG	2 (1 RH)	90 (44 RH)	98 (52 RH)	1,200	64	2	600 ms	72
Participant 10	59	F	AMC	sEEG	2 (0 RH)	113 (0 RH)	124 (0 RH)	1,200	64	2	600 ms	72
Participant 11	23	M	AMC	ECoG	58 (33 RH)	209 (110 RH)	216 (110 RH)	1,200	64	2	600 ms	72
Participant 12	39	M	AMC	sEEG	5 (5 RH)	112 (112 RH)	128 (128 RH)	1,200	64	2	600 ms	72
Participant 13	29	M	AMC	ECoG	9 (0 RH)	126 (0 RH)	132 (0 RH)	1,200	64	2	600 ms	72
Participant 14	36	M	AMC	sEEG	3 (2 RH)	169 (84 RH)	184 (90 RH)	1,200	64	2	600 ms	72
Participant 15	25	M	BJH	sEEG	19 (16 RH)	183 (93 RH)	183 (93 RH)	1,000	64	2	600 ms	72
Participant 16	38	M	BJH	sEEG	49 (15 RH)	169 (72 RH)	224 (112 RH)	1,000	64	2	600 ms	72
Participant 17	31	F	BJH	sEEG	17 (0 RH)	228 (30 RH)	228 (30 RH)	1,000	64	2	600 ms	72
Participant 18	40	M	BJH	sEEG	35 (5 RH)	137 (11 RH)	192 (14 RH)	1,000	64	2	600 ms	72
Participant 19	66	M	BJH	sEEG	32 (1 RH)	210 (13 RH)	234 (16 RH)	2,000	64	2	600 ms	72
Participant 20	24	M	BJH	sEEG	7 (0 RH)	156 (30 RH)	218 (30 RH)	2,000	64	2	600 ms	72
Participant 21	39	M	MCJ	sEEG	11 (1 RH)	108 (45 RH)	109 (45 RH)	1,200	64	1	600 ms	36
Participant 22	21	F	SLCH	sEEG	81 (81 RH)	176 (176 RH)	186 (186 RH)	2,000	64	2	600 ms	72

Data were collected at four sites: Albany Medical Center (AMC), Barnes-Jewish Hospital (BJH), Mayo Clinic Jacksonville (MCJ) and St Louis Children's Hospital (SLCH). 'Total electrodes' excludes reference electrodes, ground electrodes, microphone electrodes, trigger electrodes, skull EEG electrodes and EKG electrodes; and 'Total clean electrodes' excludes electrodes with significant line noise, significant interictal discharges, or large visual artefacts identified through manual inspection. 'Elec. per amp.' is the number of electrodes per amplifier. 'Pres. rate (per word)' is the duration of presentation for each single word or non-word.

mostly in Clusters 1 and 2, exhibited high partial correlations with another cluster's medoid (that is, a 'mixed' response profile). Visual inspection of these response profiles (Extended Data Fig. 4c,d; see OSF[37]) revealed that these electrodes displayed a blend of Cluster 1 and Cluster 2 response characteristics. The existence of mixture electrodes primarily between Clusters 1 and 2 is in line with the generally high correlation between their medoids (0.68 between Clusters 1 and 2 medoids vs 0.21 between Clusters 1 and 3, and 0.24 between Clusters 2 and 3; Fig. 3a).

Profiles reflect different temporal receptive windows

The temporal dynamics of the neural responses across clusters suggested that the observed differences in the response profiles may reflect different 'temporal receptive windows' (TRWs). TRWs are a temporal equivalent of spatial receptive fields that corresponds to the amount of the preceding temporal context that affects the processing of the current input (for example, refs. 32,40,41). In particular, a neural population that only processes information over the span of a single word should exhibit visible evoked responses at the rate of stimulus presentation, reflecting the momentary stimulus-related fluctuations. On the other hand, a neural population that processes information over spans of multiple words should exhibit a response that reflects a more smoothed version of the stimulus train, with no momentary stimulus-related fluctuations. As described in 'Language electrodes exhibit distinct response profiles', the three clusters differed in their degree of locking to the individual word onsets. Cluster 3 showed the strongest locking, followed by Cluster 2, with Cluster 1 showing the weakest amount of locking (Fig. 3c), consistent with a decreasing TRW size from Cluster 1 to 2 to 3. Moreover, a neural population that only processes information over the span of roughly a single word (or less) should show little sensitivity to whether nearby words can be composed into phrases. This is the pattern we saw for electrodes in Cluster 3 (Fig. 3d): these electrodes did not reliably discriminate between the Sentence and Word-list conditions. In contrast, a population that processes information over spans of multiple words should show sensitivity to the composability of nearby words, and thus should strongly discriminate between sentences and word lists. This is the pattern we saw for electrodes in Clusters 1 and 2, with Cluster 1 electrodes showing earlier and stronger discrimination (Fig. 3d). Note that this greater difference between the Sentence and Word-list conditions for longer-TRW neural populations is presumably due to the fact that linguistic differences between these two conditions become more pronounced for longer word sequences (for example, see Extended Data Fig. 5 for evidence from *n*-gram frequency counts).

To formally test whether the clusters indeed differ in the size of their TRWs, we constructed a toy model wherein we convolved a simplified stimulus train with response functions (Gaussian-based 'kernels') of varying widths (TRW sizes denoted as σ; Fig. 4a, see Methods for model assumptions and implementational details). The resulting simulated responses exhibited striking visual similarity to the observed response patterns (Fig. 4a). We then computed, for every electrode, a correlation between each simulated response and the observed response, and we selected the σ value that yielded the highest correlation (Fig. 4b,c and Methods). The estimated TRW sizes showed a clear pattern of decrease from Clusters 1 to 2 to 3; the average σ values per cluster were ~6, ~4 and ~1 words for Clusters 1, 2 and 3, respectively ($P < 0.0001$ comparing TRWs across all pairs of clusters, evaluated with a linear mixed-effects (LME) model; Methods and Fig. 4b,c, see Supplementary Table 5 for a complete description of the statistical details and results). To evaluate the robustness of this result, we repeated the TRW fitting procedure using other kernel shapes and confirmed that the relative sizes of the TRWs of the three clusters did not depend on the specific choice of kernel shape (Extended Data Fig. 6). Furthermore, the estimated values of σ in number of words (as reported above) appear to be invariant to the stimulus presentation rate, which suggests that the TRW of language-responsive electrodes is information-, not time-, dependent (see Supplementary Tables 6 and 7 for complete statistical results). However, this rate invariance should be investigated further in future work given the small number of participants in each presentation rate group ($n = 3$) and, correspondingly, the low statistical power.

https://doi.org/10.1038/s41562-024-01944-2

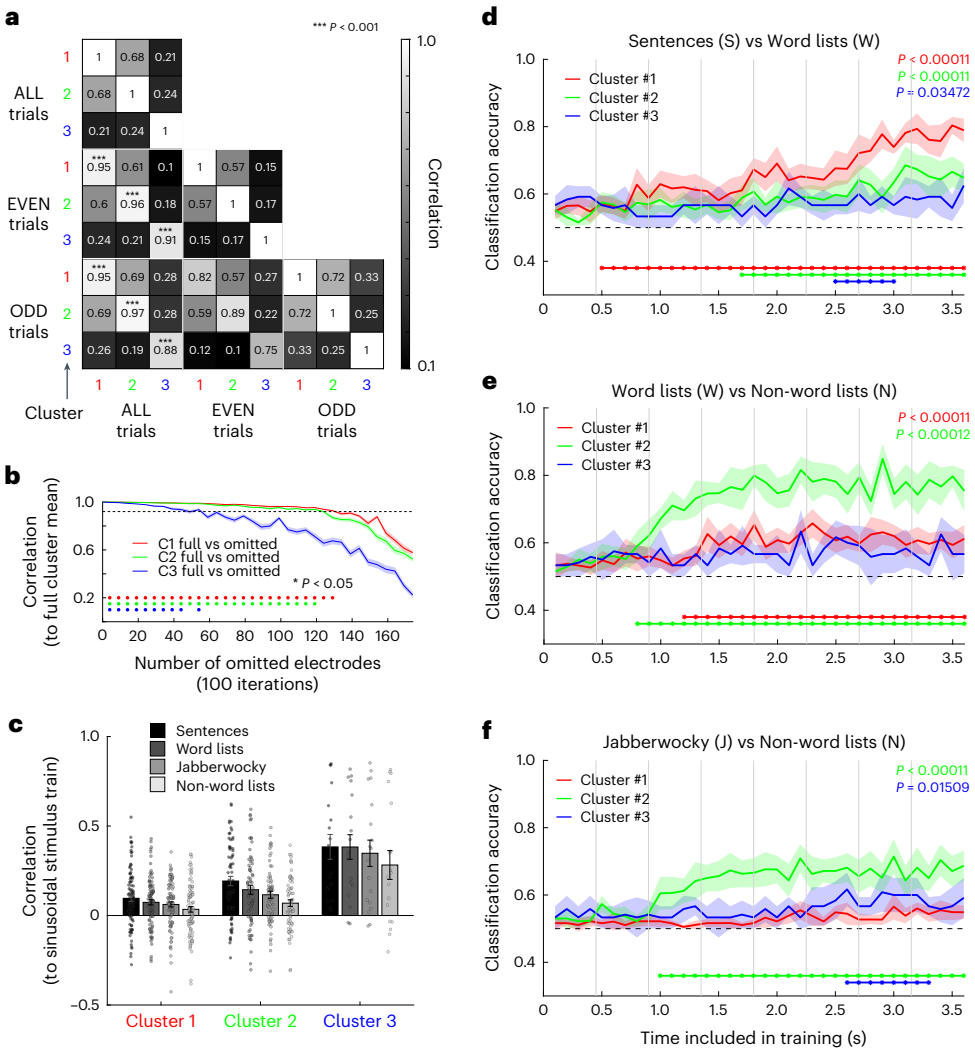


Fig. 3 | Evaluation of Dataset 1 clusters. a, Comparison of clusters from all trials (top three rows) versus only even (middle three rows) or odd (bottom three rows) trials. Clusters that emerged using only odd or even trials were highly similar to the clusters that emerged when all trials were used ($P < 0.001$; evaluated with a one-sided permutation test, $n = 1{,}000$ permutations; Methods). **b**, Robustness of clusters to electrode omission. Random subsets of electrodes were removed in increments of 5 (Methods). Similarity of cluster centres was computed between the clusters that emerged when all electrodes were used versus when random subsets of electrodes were removed. Stars reflect significant similarity with the full dataset (using a P threshold of 0.05; evaluated with a one-sided permutation test, $n = 1{,}000$ permutations; Methods). Shaded regions reflect s.e.m. over randomly sampled subsets of electrodes. Cluster 3 was driven the most by individual electrodes relative to Clusters 1 and 2. **c**, Correlation of fitted sinusoidal function with timecourse of electrodes (averaged across trials) by cluster and by condition (Methods). Error bars reflect s.e.m. over electrodes ($n = 92$, 67 and 18 electrodes for Clusters 1, 2 and 3, respectively). Data points represent individual electrodes. Electrodes in Cluster 3 were the most locked to word/non-word presentation, whereas electrodes in Cluster 1 were the least locked to word/non-word presentation. There was a significant main effect for cluster ($P < 0.05$) but not for condition (two-sided ANOVA for linear mixed-effects models; Methods, see Supplementary Table 1a,b for complete statistical results). The observed qualitative between-condition differences could be due to generally greater engagement of these neural populations with more language-like stimuli. **d**–**f**, Classifier performance by cluster as a function of the amount of timecourse included in training (Methods). A binary logistic classifier was trained to discriminate the Sentence (S) and Word-list (W) conditions (**d**), Word-list (W) and Non-word-list (N) conditions (**e**), and Jabberwocky (J) and Non-word-list (N) conditions (**f**). Significance stars at the bottom (coloured by cluster) reflect discriminability of conditions above chance level ($P < 0.05$, evaluated as a one-sided cluster statistic against a null distribution from permuted labels, $n = 1{,}000$ permutations; this statistical test accounts for multiple comparisons and for the autocorrelational structure of the signal; Methods and refs. 38,39). Shaded regions reflect s.e.m. across the 10 folds of the cross-validated classifier. The dashed black line reflects chance performance (0.5).

Clusters 1 and 2 are interleaved, Cluster 3 shows posterior bias

We tested for differences in the anatomical distribution of the electrodes that belong to the 3 clusters in Dataset 1. We excluded from this analysis right-hemisphere (RH) electrodes because only 4 RH electrodes passed the language selectivity criterion (S > N). We focused on the y (posterior-anterior) and z (inferior-superior) directions in the MNI coordinate space within the left hemisphere. Electrodes in both Clusters 1 and 2 were distributed across the temporal and frontal language regions (Fig. 5). When examining all electrodes together, or focusing on only the frontal or only the temporal electrodes, the MNI coordinates of electrodes in Clusters 1 and 2 did not significantly differ in either of the two tested directions ($P > 0.05$, evaluated with an LME model; Methods and Fig. 5c,d, see Supplementary Table 2a for complete statistical results). However, when weighting the electrodes by their reliability in the LME model, electrodes in Cluster 1 fell more anteriorly and inferiorly relative to electrodes in Cluster 2 ($P < 0.05$, evaluated with an LME model; Methods, see Supplementary Table 2b for complete statistical results). Electrodes in Cluster 3 were located

https://doi.org/10.1038/s41562-024-01944-2

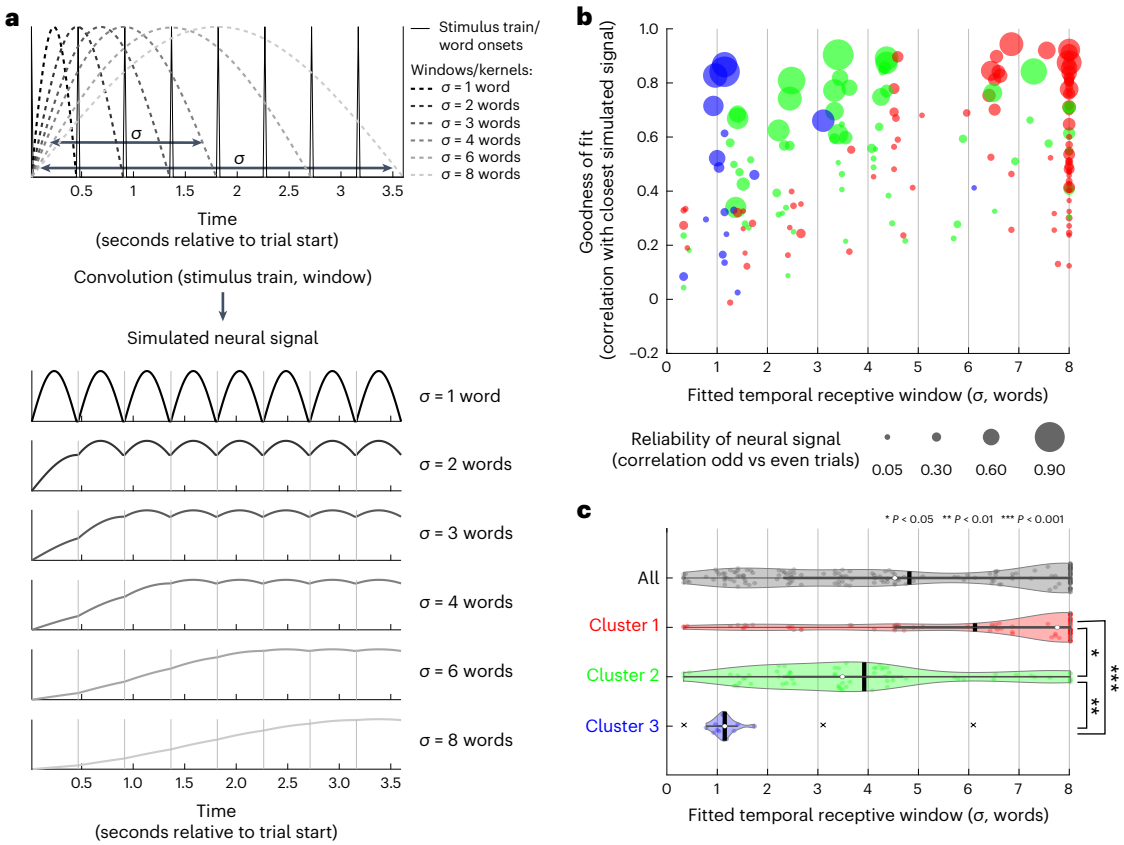


Fig. 4 | Estimating the size of the TRW of different electrodes. a, A toy model that simulates neural responses to the Sentence condition as a convolution of a simplified stimulus train and truncated Gaussian kernels of varying widths. Top: simplified stimulus train where peaks indicate a word/non-word onset and sample kernels correspond to varying temporal receptive window sizes (σ). The kernels were constructed from Gaussian curves with a standard deviation of $\sigma/2$ truncated at ±1 s.d. (capturing 2/3 of the area under the Gaussian; Methods) and normalized to a minimum of 0 and a maximum of 1. Bottom: the resulting simulated neural signals for sample kernel widths, normalized to a minimum of 0 and a maximum of 1. **b**, Best TRW fit for all electrodes (each dot represents a single electrode) coloured by cluster and sized by the reliability of the neural signal as estimated by correlating responses to odd and even trials (Fig. 1d). The goodness of fit, or correlation between the simulated and observed neural signal (Sentence condition only), is shown on the *y* axis. **c**, Estimated TRW sizes across all electrodes (grey) and per cluster (red, green and blue representing Clusters 1, 2 and 3, respectively). Within each horizontal violin plot, single dots represent single electrodes, black vertical lines correspond to the mean window size and the white dots correspond to the median. Horizontal thin black boxes represent the lower and upper quartiles, 'x' marks (present in Cluster 3 only) indicate outlier electrodes (more than 1.5 interquartile ranges above the upper quartile or less than 1.5 interquartile ranges below the lower quartile) and the violin body is plotted within the data range after excluding outliers. Significance was evaluated with an LME model (comparing estimate values, two-sided ANOVA for LME; Methods, see Supplementary Table 5 for exact *P* values). Together, **b** and **c** show that the clusters varied in the size of their TRWs, from a relatively long TRW (Cluster 1) to a relatively short one (Cluster 3).

significantly more posteriorly than those in Clusters 1 and 2 (lower *y*-coordinate values, both Clusters 3 vs 1 and Clusters 3 vs 2, $P < 0.0001$; Methods and Fig. 5c, see Supplementary Table 2a for complete statistical results).

To complement this analysis, we visualized the anatomical distribution of electrodes in two additional ways. First, we visualized all language-responsive electrodes by their partial correlations to each of the cluster medoids (Extended Data Fig. 4e). This approach does not enforce a categorical grouping into clusters, potentially allowing for more subtle response gradients. However, this analysis revealed a similar picture: Cluster-1- and Cluster-2-like responses were present throughout frontal and temporal areas, whereas Cluster-3-like responses were localized to the posterior superior temporal gyrus. Second, we examined the distribution of electrodes by their fitted TRW (Extended Data Fig. 5f). This visualization exhibited a gross anatomical trend of TRWs increasing from posterior to anterior regions; however, there remained a substantial local mosaic pattern, with long-TRW electrodes present in posterior temporal areas and short-TRW electrodes present in anterior temporal and frontal areas as well.

Clusters 1 and 3 replicate in Dataset 2, Cluster 2 partly replicates

We asked whether the same clusters would emerge in a second, independent dataset with new participants and different linguistic materials (Dataset 2; 16 participants; 362 language-responsive electrodes; mostly depth electrodes; Fig. 6a, Methods and Table 2). Participants in Dataset 2 only saw two of the four conditions presented to participants in Dataset 1 (sentences (S) and non-word lists (N), but not word-lists (W) or Jabberwocky sentences (J)); therefore, we started by reclustering the electrodes from Dataset 1 using only the responses to the S and N conditions to allow for direct comparisons with Dataset 2.

The Dataset 1 cluster averages, when only the S and N conditions were used, exhibited a strong qualitative similarity to those of the clusters derived using the data from all four conditions (Extended Data Fig. 7). Approximately 80% of electrodes in Dataset 1 were assigned to the same cluster regardless of whether they were clustered using all four or just the two conditions (the S+N clusters were 'matched' to the S+W+J+N clusters by highest correlation, Methods). However, Cluster 2 was less robust to electrode loss than Clusters 1 and 3 (compare the green curve in Fig. 3b to the green curve in Extended Data Fig. 7g).

https://doi.org/10.1038/s41562-024-01944-2

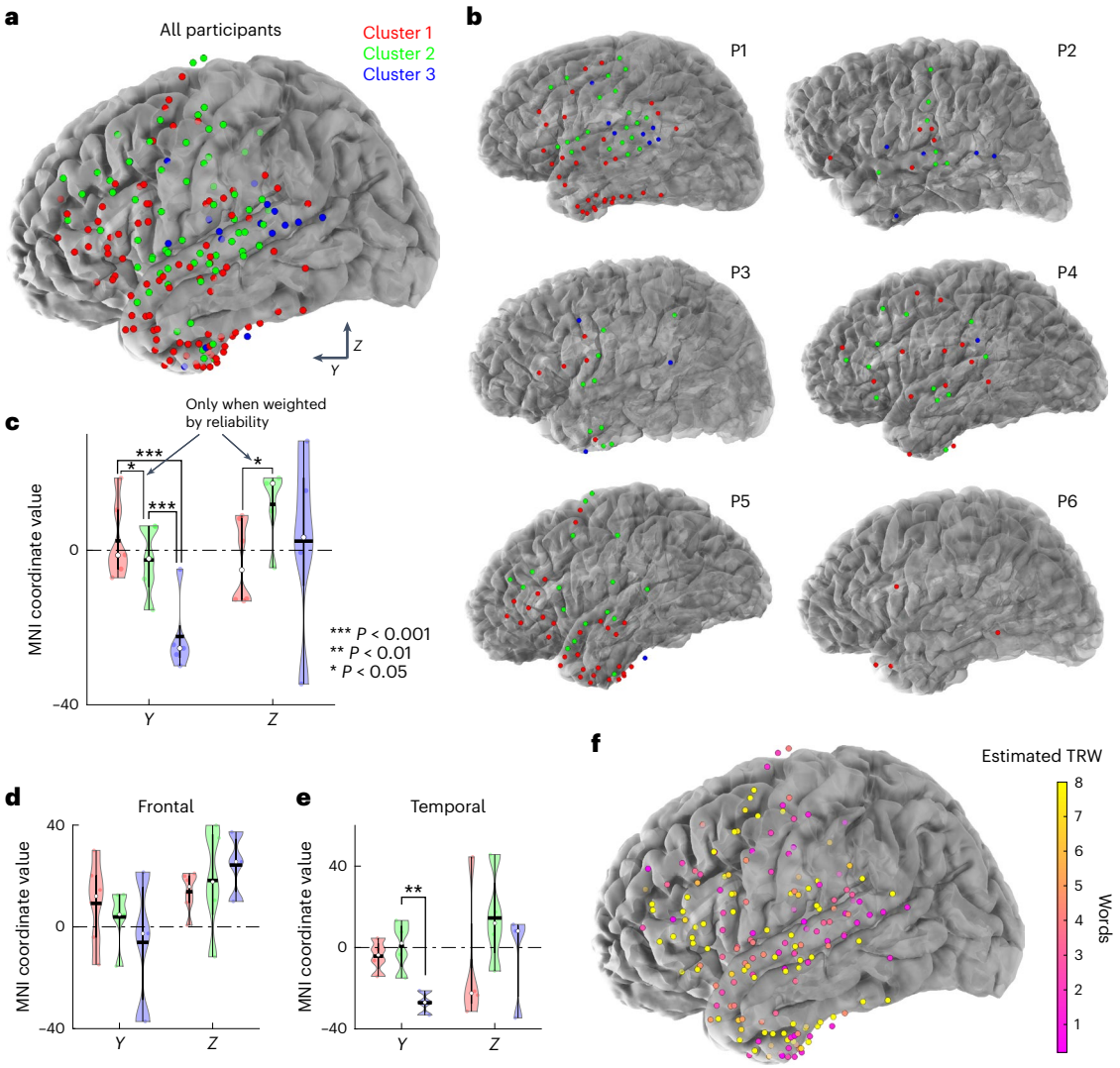


Fig. 5 | Anatomical distribution of the clusters in Dataset 1. a, Anatomical distribution of language-responsive electrodes in Dataset 1 across all participants in MNI space, coloured by cluster. **b**, Anatomical distribution of language-responsive electrodes in participant-specific space, coloured by cluster. **c**–**e**, Violin plots of MNI coordinate values for the 3 clusters, where each point represents the mean coordinate value across electrodes for a given participant and cluster. Individual dots represent individual participants ($n = 6$). The mean across participants is plotted with a black horizontal line and the median is shown with a white circle. Vertical thin black boxes represent the lower and upper quartiles. Significance was evaluated with a linear mixed-effects (LME) model (comparing estimate values, two-sided ANOVA for LME models; Methods, for exact P values see Supplementary Table 2a–d). Cluster 3 exhibited a posterior bias (more negative y-coordinate) relative to Clusters 1 and 2 when modelled using all language electrodes ($P < 0.001$, **c**). This trend was also evident when examining only the frontal (**d**) or only the temporal (**e**) electrodes separately, but the difference only reaches significance for the temporal electrodes ($P < 0.01$). **f**, Anatomical distribution of electrodes in Dataset 1 coloured by their estimated TRW (Fig. 4). There was a slight trend of increasing TRW size from posterior to anterior regions but with considerable local heterogeneity.

This finding suggests that responses to the Word-list and Jabberwocky conditions are especially important for differentiating Cluster 2 from the other response profiles, presumably because these conditions pattern differently for Clusters 1 and 2.

We next clustered the electrodes in Dataset 2 using the same approach as for Dataset 1. The optimal number of clusters in Dataset 2 was $k = 2$ based on the elbow method, and the resulting clusters were visually similar to Clusters 1 and 3 from Dataset 1 ($n = 211$ electrodes and $P < 0.001$ for the cluster resembling Cluster 3 in Dataset 1; $n = 151$ electrodes and $P = 0.061$ for the cluster resembling Cluster 1 in Dataset 1; one-sided permutation test, $n = 1{,}000$ permutations; Methods, see OSF[37]; note that this permutation test is especially conservative with only two experimental conditions and when $k = 2$). We also performed a version of clustering Dataset 2 enforcing $k = 3$ to test whether a Cluster-2-like response would emerge (Fig. 6). The same two cluster centres as in the case of $k = 2$ were again apparent and showed reliable similarity to Clusters 1 and 3 in Dataset 1 ($n = 172$ electrodes and $P < 0.001$ for the cluster resembling Cluster 1 in Dataset 1; $n = 81$ electrodes and $P = 0.023$ for the cluster resembling Cluster 3 in Dataset 1; one-sided permutation test, $n = 1{,}000$ permutations; Methods and Fig. 6g,i). The third cluster qualitatively resembled some aspects of Cluster 2 from Dataset 1 (Fig. 6g), but the resemblance was not statistically reliable ($n = 109$ electrodes, $P = 0.732$, one-sided permutation test, $n = 1{,}000$ permutations; Methods).

As another, less stringent, test of whether Cluster 2 responses were present in Dataset 2, we assigned each electrode in Dataset 2 to a 'group' on the basis of their highest correlation with the average response profiles from Dataset 1, in a 'winner-take-all' approach (Extended Data Fig. 8). A substantial number of electrodes in Dataset 2 ($n = 95$ of the total of $n = 362$) were most correlated with Cluster 2 from Dataset 1

https://doi.org/10.1038/s41562-024-01944-2

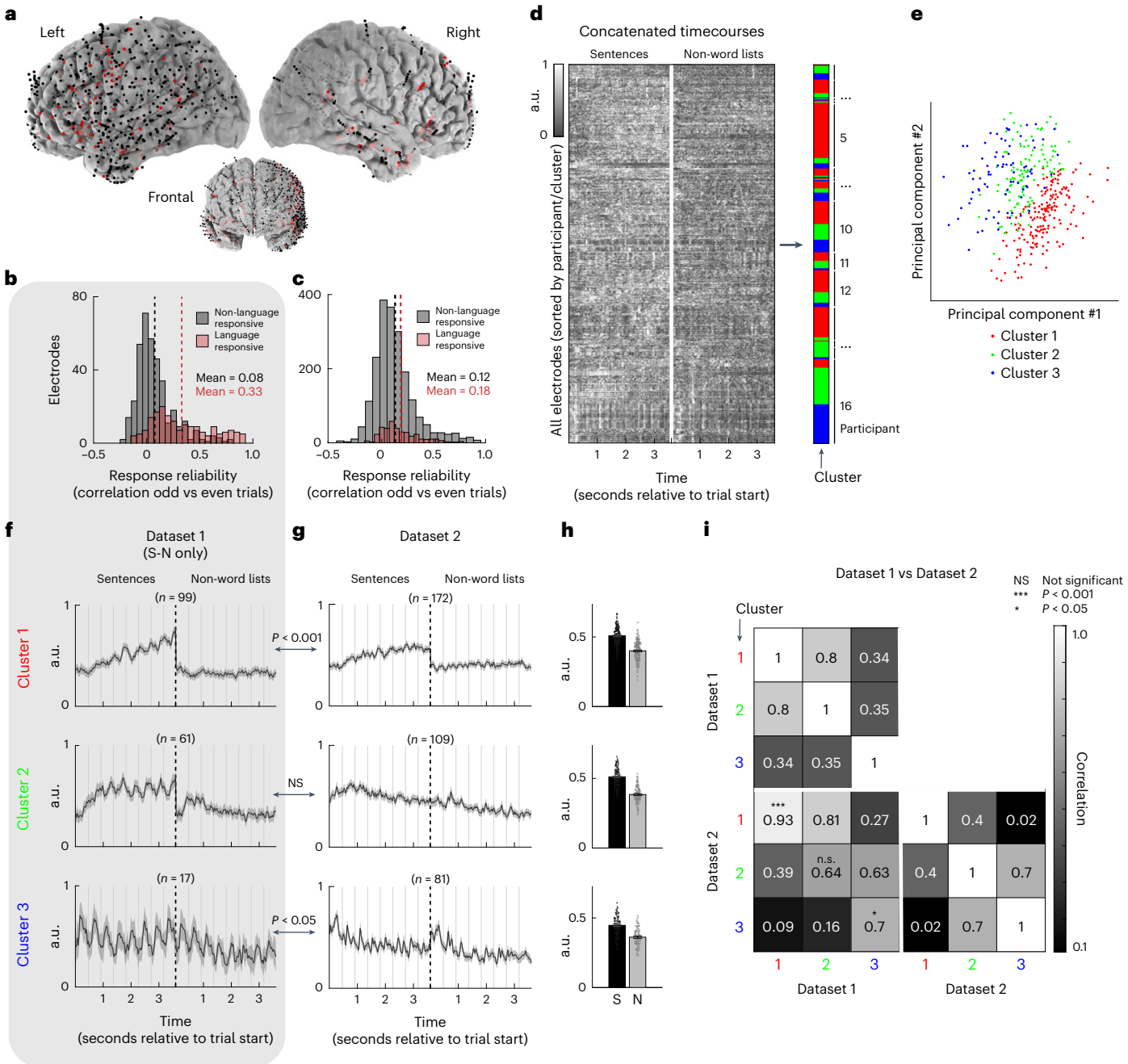


Fig. 6 | Dataset 2 *k*-medoids clustering with $k = 3$. **a**, The locations of language-responsive ($n = 362$, red; Methods) and non-language-responsive ($n = 2{,}017$, black) electrodes across the 16 participants in Dataset 2 (both surface and depth electrodes were implanted). Language-responsive electrodes were found across the cortex, in both the left and right hemispheres (Table 2). **b**,**c**, Response reliability as estimated by correlating responses to odd and even trials for language-responsive and non-language-responsive electrodes (as in Fig. 1d). Language-responsive electrodes exhibit more reliable responses to linguistic stimuli than non-language-responsive electrodes for both Dataset 1 (when taking into account the Sentence and Non-word-list conditions only, **b**) and Dataset 2 (**c**); however, the responses of language electrodes were less reliable in Dataset 2 than in Dataset 1. **d**, Clustering mean electrode responses (concatenated responses to sentences and non-word lists) in Dataset 2 using *k*-medoids ($k = 3$) with a correlation-based distance. Shading of the data matrix reflects normalized high-gamma power (70–150 Hz). **e**, Electrodes visualized on their first two principal components, coloured by cluster. **f**,**g**, Average timecourse by cluster from Dataset 1 when using only the Sentence and Non-word-list conditions (**f**; see Extended Data Fig. 7; $n = 99$, 61 and 17 electrodes for Clusters 1, 2 and 3, respectively) and from Dataset 2 (**g**, $n = 172$, 109 and 81 electrodes for Clusters 1, 2 and 3, respectively). Shaded areas around the signal reflect the 99% confidence interval over electrodes. **h**, Mean condition responses by cluster in Dataset 2. Error bars reflect s.e.m. over electrodes ($n = 172$, 109 and 81 electrodes for Clusters 1, 2 and 3, respectively, as in **g**). As with Dataset 1, after averaging across time, response profiles were not as distinct by cluster, underscoring the importance of temporal information in elucidating this grouping of electrodes. **i**, Evaluation of clusters from Dataset 1 (clustering with the Sentence and Non-word-list conditions only) against clusters from Dataset 2. Clusters 1 and 3 from Dataset 1 replicated in Dataset 2 ($P < 0.001$ and $P = 0.023$, respectively; one-sided permutation test; Methods). Although Cluster 2 demonstrated some qualitative similarity across the two datasets, this similarity was not statistically reliable ($P = 0.732$, one-sided permutation test; Methods).

(while $n = 142$ electrodes were most correlated with Cluster 1 and $n = 125$ electrodes were most correlated with Cluster 3). This analysis indicates that Cluster-2-like responses are indeed present in Dataset 2, even though they did not reliably emerge through the data-driven clustering approach. The lower robustness of the Cluster-2-like responses in Dataset 2 could in part be attributable to the lower split-half reliability of Dataset 2 compared with Dataset 1 (compare Fig. 6b vs c), as well as the sparser spatial coverage due to the prevalence of depth electrodes (Fig. 6a). For completeness, an analysis of the anatomical trends in Dataset 2 is presented in Extended Data Fig. 9 (see Supplementary Tables 3 and 4 for a complete statistical comparison of the anatomical locations of the two statistically reliable clusters in Dataset 2).

Finally, we estimated the TRW size (as in 'Profiles reflect different temporal receptive windows') for each electrode in Dataset 2 (Extended Data Fig. 10). Clusters 1 and 3 (the two clusters that consistently replicated from Dataset 1) were best described by TRWs of ~4.5 and ~1 words, respectively (Extended Data Fig. 10a,b), similar to the TRW sizes observed for those clusters in Dataset 1. The TRW of Cluster 2 did not significantly differ from that of Cluster 3 when relying on the electrode assignments from the clustering algorithm with $k = 3$ (where Cluster 2 did not replicate; Methods, Fig. 6, Extended Data Fig. 10b and Supplementary Table 8). However, using the winner-take-all approach (where a more Cluster-2-like response was observed; Extended Data Figs. 8 and 10d), the TRW of Group 2 was ~2.1 words, which significantly differed from that of Groups 1 and 3 ($P < 0.01$ comparing TRWs across all pairs of groups, evaluated using an LME model; Methods and Extended Data Fig. 10c,d, see Supplementary Table 9 for complete statistical results) and was similar to the TRW of Cluster 2 from Dataset 1.

Discussion

The nature of the neural computations that support our ability to extract meaning from linguistic input remains an important open question in the field of language research. Here we leveraged the high temporal and spatial resolution of human intracranial recordings to probe the fine temporal dynamics and the spatial distribution of language-responsive neural populations. We uncovered three temporal profiles of response during the processing of sentences and linguistically degraded conditions such as lists of words or non-words. We suggest that these profiles differ in the size of their TRW—the amount of temporal context that affects the neural processing of the current input. Further, we found that electrodes with distinct response profiles manifest in a scattered spatial distribution across both frontal and temporal cortices. Below, we contextualize these results with respect to previous empirical work and discuss their implications for our understanding of human language processing.

Three types of language-responsive neural populations

In the present study, we used a clustering approach to group neural populations (as measured by intracranial macroelectrodes; note that when we write that 'electrodes' exhibit a response, we are referring to the 'neural populations' that the electrodes are measuring) by their responses to four types of language stimuli: sentences (S), lists of unconnected words (word lists, W), Jabberwocky sentences (where content words are replaced with pronounceable non-words; J) and lists of unconnected non-words (N). We uncovered three dominant response profiles ('clusters') that differed in the presence and timing of the increase (buildup) of neural activity over the course of a sentence, the degree of locking to individual word/non-word onsets and the overall magnitude of response to the linguistically degraded conditions (W, J and N). Within each cluster, individual electrodes exhibited highly similar responses, with a small number of electrodes displaying a mixed response between Clusters 1 and 2. Finally, we found evidence for each of the three response profiles in an independent dataset that only included two of the four linguistic conditions (sentences and non-word lists), although Clusters 1 and 3 were more robustly replicated. Importantly, because we had restricted our analyses to electrodes that show a functional signature of the language network (a stronger overall response during the processing of structured and meaningful language stimuli—sentences—than during the processing of perceptually similar but meaningless and unstructured stimuli—non-word lists; ref. 1), these findings provide evidence for 'functional heterogeneity within the language network' proper, rather than between the language areas and nearby functionally distinct brain regions, such as speech areas[42,43] or higher-level cognitive networks[44,45] (see ref. 12 for discussion).

The experimental design adopted in the current study has traditionally been used as a way to tease apart neural responses to word meanings (present in sentences and word lists, but not in Jabberwocky sentences and non-word lists) and syntactic structure (present in sentences and, under some views of syntax, in Jabberwocky sentences, but not in word/non-word lists; refs. 1,27,35). As measured with fMRI, all areas of the language network show sensitivity to both word meanings and syntactic structure: the response is strongest to sentences, lower to word lists and Jabberwocky sentences, and lowest to non-word lists[1,2,30,36] (see refs. 23,22 for evidence against the lexical/syntactic dissociation from other paradigms; see ref. 46 for earlier arguments and evidence). Using a similar design in an intracranial recording study, ref. 27 replicated this overall pattern of response and also reported a temporal profile—present in a subset of electrodes—whereby high-gamma power builds up across words over the course of a sentence but not in other conditions (replicated in refs. 28–30). They interpreted this build-up effect as indexing the process of constructing a sentence-level meaning.

Here we investigated the temporal profiles of language-responsive electrodes more comprehensively. By leveraging the fine-grained temporal information in the signal (that is, considering the full timecourses instead of averaging high-gamma power in each word/non-word as in ref. 27), we found that the build-up effect reported in ref. 27 represents a mix of functionally distinct populations. The timecourse of response to the Sentence condition in ref. 27 is most similar to that in Cluster 1 here. However, a reliable sentences > word lists > Jabberwocky sentences > non-word lists profile in ref. 27 suggests a contribution from Cluster 2 neural populations. As such, our analyses identify two functionally distinct build-up profiles and additionally uncover a third profile, which does not show buildup of activity over time, and we replicated these results in a new, larger dataset with a different set of language materials (Dataset 2). Importantly, here we show that despite strong integration between lexical and syntactic processing, neural populations within the language network 'do' differ functionally, although along a different dimension—the temporal scale of information integration.

Distinct temporal receptive windows in the language network

A TRW denotes the amount of the preceding context that a given neural unit integrates over[32,40,41]. Previous studies have demonstrated that cortical neural activity is organized into a hierarchy of timescales, wherein information over tens to hundreds of milliseconds is encoded by sensory cortical areas, and information over many seconds is encoded by higher-order areas[47–49]. Past fMRI studies have shown that the TRW of the language network falls somewhere between a word and a short sentence[32–34,36,50–52], although some work has suggested that language regions are, at least to some degree, sensitive to sublexical regularities[25,53]. Using a simple instantiation of an information processing system—with one (interpretable) free parameter: the length of past stimulus context—we estimated the TRW of different language-responsive neural populations. On the basis of this analysis, we argue that our observed 'response profiles differ in their timescale of information processing', from sublexical units and single words (Cluster 3) to short phrases (Cluster 2) to longer phrases/sentences (Cluster 1).

Do the observed response profiles reflect categorically distinct clusters that integrate information over different timescales, or is

the underlying structure of language-selective responses in the brain best described by a continuum of TRWs with no sharp boundaries or groupings of response types? Although we do not rule out the possibility of a TRW continuum, our data are well explained by the grouping of responses into three categories. A few electrodes do exhibit a 'mixed' response profile, falling somewhere between the prototypical Cluster 1 and Cluster 2 responses, but this mixing could be due to these electrodes picking up activity of multiple neural populations. Recordings at a higher spatial resolution would be needed to evaluate this possibility (for example, refs. 54,55). Nevertheless, the current data suggest the existence of neural populations within the language network that are sensitive to information chunks of 'distinct and specific size'. This functional organization is presumably driven by the statistics of natural language and is probably critical for efficient extraction of meaning from language (see 'Future directions').

To estimate the TRW values, we made several simplifying assumptions that can be revisited in future studies. First, we have discussed TRWs in terms of the number of 'words'. However, natural languages vary substantially in how they package information into words[56] and the processing of a given word is highly dependent on how informative the word is in context (for example, ref. 57; for behavioural evidence, see ref. 58). As a result, TRWs may instead be bounded by the number of bits of information. Future work should evaluate multiple accounts of the units in which TRWs are measured. The second simplifying assumption we made was that TRWs are fixed in size. Much recent evidence suggests that human comprehension mechanisms can flexibly accommodate corrupt linguistic input, for example, due to speech errors (for example, refs. 59–61; see ref. 62 for a review), which may make it desirable for TRWs to be somewhat adaptable to allow for the possibility of continuously revising one's interpretation of the input. Future work should seek to understand if and how the TRW of a specific neural population can be affected by linguistic context. And third, the response function (kernel) that we used to generate the simulated signals was intentionally simple and was not designed to be fully consistent with the underlying neurophysiology (see Methods for details). A model that is more faithful to neurobiological principles may better capture the observed neural responses and such models should be explored in future work.

Finally, our toy TRW model currently does not take into account the form and content of the stimulus, as it does not use any linguistic information to generate responses. However, responses of neural populations in the language network are highly sensitive to linguistic stimulus properties. One key modulator of response strength is how well the stimulus matches natural language statistics, as evidenced by both condition-level effects (for example, sentences > word lists; ref. 1) and fine-grained preferences for particular linguistic strings[63]. A more complete model of language processing should therefore include both 'gating' of linguistic input into different lengths of effective input (defined by a neural population's TRW) and 'scaling' of the neural response by the effective input's probability. This idea—that responses of neural populations in the language network reflect the probability of linguistic inputs at variable context lengths due to their TRW—may explain why the Sentence and Word-list conditions were best discriminated by Cluster 1 populations. In particular, Cluster 1 populations have the longest TRW, and the linguistic difference between sentences and word lists becomes more apparent over longer timescales (Extended Data Fig. 5). We leave more thorough exploration of stimulus-dependent accounts of the computations carried out by the language network to future work (see 'Future directions').

The spatially distributed nature of language processing

There is a long history in language neuroscience of attempts to divide language comprehension into both temporally distinct stages and spatially distinct components. At some level, language comprehension can indeed be broken up across time and space. In particular, clear separation exists between the language-processing system[7] and both (1) lower-level perceptual areas and (2) higher-level cognitive areas (see ref. 12 for a review). During language perception, the lower-level perceptual areas, such as the speech perception area[42,43,64] and the visual word-form area[65], process information 'earlier' than, and probably provide input to, the language network. And higher-level cognitive areas, such as the areas of the Default network[66] or the Theory of Mind network[67], process information 'later' than, and probably receive input from, the language network. These latter areas plausibly carry out further processing on the meaning representations extracted from language, including connecting those meaning representations across long spans of time[32,68]. However, discovering spatial subdivisions 'within' the language-selective network proper has proven challenging[1,22,23,33,36].

The current work demonstrates that there exist functional differences within the language network, but functionally distinct populations do not seem to exhibit strong spatial clustering and are instead distributed in an interleaved fashion across the language network. The latter explains why most past fMRI work could not reveal this functional heterogeneity (cf. ref. 35 for implied functional heterogeneity based on multivariate patterns of fMRI response; and see ref. 34 for evidence of voxel-level heterogeneity with respect to TRWs as discovered in an encoding approach with artificial neural network language models). This architectural design makes it possible for each area of the network to have access to information at different timescales, which probably makes language processing efficient and robust. A clear exception in our data is the concentration of Cluster 3 (shortest-TRW) electrodes in the posterior superior temporal gyrus, which may suggest that this area serves a unique computational role within the language network (see refs. 69,36 for other recent evidence of the special role of this area); however, we cannot rule out the possibility that these electrodes are picking up some activity from the nearby speech areas[42]. We also acknowledge that a macroscale organization could become more evident with more participants and a more systematic coverage of the frontal and temporal cortex.

Future directions

The current findings lay the foundation for several exciting future research avenues. First, the size of a neural unit's TRW should determine its sensitivity to different linguistic features. As noted above, one limitation of the current investigation is the focus on condition-level differences, rather than trying to explain fine-grained responses to individual linguistic items. The reason for this choice is twofold. To start, the current linguistic materials were not constructed with the goal of investigating linguistic (for example, lexical and syntactic) features; to make the materials easy to process for diverse populations, the sentences were constructed to be short and to use common structures and words, which limits the range of variability to be explored. In addition, we did not observe reliable stimulus-related activity (beyond the level of conditions; see OSF[37]). However, the TRW-based framework makes clear predictions that can be evaluated in future work. For example, short-TRW populations should show greater sensitivity to lexical features, such as word frequencies, whereas longer-TRW populations should be more sensitive to linguistic features at longer timescales, such as higher-order *n*-gram frequencies and syntactic-structure-related features. Because many linguistic features are strongly intercorrelated in naturalistic language materials[70,71] (see OSF[37] for evidence of intercorrelation of linguistic features in the current stimuli), evaluating these predictions will require constructing materials that are specifically designed to best dissociate different linguistic dimensions.

Second, artificial neural network (ANN) language models, which have proven to be powerful tools for understanding the human language system[31,34,72] (see ref. 73 for a review), could be leveraged to gain insights into the constraints on the language processing architecture. For example, do successful language architectures require particular proportions of units with different TRWs or particular distributions of

such units within and/or across model layers? In Dataset 1, we found the fewest electrodes belonging to Cluster 3 (shortest TRW), more electrodes belonging to Cluster 2 (intermediate TRW) and the majority of electrodes belonging to Cluster 1 (longest TRW). These proportions align with the idea that compositional semantic space is highly multidimensional, but word-form information can be represented in a relatively low-dimensional space[74]. However, the proportions can also be affected by biases in where intracranial electrodes tend to be implanted, so investigating these questions in ANNs, where we can probe all units in the network[75] and have the freedom to alter the architecture in various ways[34], may yield insights that cannot be gained from human brains at least with the current experimental tools available.

And third, we have here focused on language comprehension. However, the same language network also supports language production[76,77]. Whether the TRW-based organization discovered here in a language comprehension task applies to language production, given that utterance planning is known to unfold at multiple scales[78], remains to be determined.

In conclusion, across two intracranial-recording datasets, we here demonstrate the existence of functionally distinct neural populations within the fronto-temporal language-selective network proper, opening the door to investigations of how these populations work together to accomplish the incredible feats of language comprehension and production.

Methods

Participants

Dataset 1 (also used in ref. 27): 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 epileptogenic zones and to delineate them from eloquent cortical areas before brain resection. Patients with a verbal IQ score >70, as defined by the Wechsler Abbreviated Scale of Intelligence-Second Edition (WASI-II, ref. 79), and general verbal proficiency, as qualitatively evaluated by the experimenters collecting the data, were eligible to participate in the study. The administration of the task was prioritized in patients with left hemisphere frontal and temporal coverage. All participants gave informed written consent to participate in the study, which was approved by the Institutional Review Board of Albany Medical Center (protocol number #2061). The participants were not compensated for their participation. One further participant was tested but excluded from analyses because of difficulties in performing the task (that is, 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 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). Patients with a verbal IQ score >70, as defined by the Wechsler Abbreviated Scale of Intelligence-Second Edition (WASI-II,[79]), and general verbal proficiency, as qualitatively evaluated by the experimenters collecting the data, were eligible to participate in the study. All participants gave informed written consent to participate in the study, which was approved by the Institutional Review Board at each relevant site (protocols #2061 (AMC), #18-011810 (MCJ) and #201102222 (BJH, SLCH)). The participants were not compensated for their participation. Two further participants were tested but excluded from analyses due to the lack of any language-responsive electrodes (see 'Language-responsive electrode selection').

Data collection

Dataset 1: The implanted electrode grids consisted of platinum-iridium electrodes that were 4 mm in diameter (2.3–3 mm exposed) and spaced with an inter-electrode distance of 0.6 or 1 cm. The total numbers of implanted grid/strip electrodes were 120, 128, 98, 134, 98 and 36 for the 6 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 1 cm (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; s.d. = 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 (v.3.6, ref. 80).

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[81,82]. Electrode locations were then transformed to MNI space within VERA via non-linear co-registration of the participants' skull-stripped anatomical scan and the skull-stripped MNI152 Freesurfer template using ANTs[83].

Preprocessing and extraction of signal envelope

Neural recordings were collected and saved in separate data files by run (see 'Experiment', and Tables 1 and 2), and all preprocessing procedures were applied 'within' data files to avoid inducing artefacts 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 were excluded from analysis: (1) ground, (2) reference and (3) those that were not ECoG or sEEG contacts (for example, microphone electrodes, trigger electrodes, scalp electroencephalography (EEG) electrodes, EKG electrodes), as well as (4) those with significant line noise, defined as electrodes with line noise greater than 5 s.d. above other electrodes, (5) those with large artefacts identified through visual inspection and, for all but four participants, (6) those that had a significant number of interictal discharges identified using an automated procedure[84]. For 4 participants (P3 in Dataset 1 and P15, P17 and P21 in Dataset 2), electrodes that were identified as having a significant number of interictal discharges were not excluded from analyses because more than 1/3 of each of these participants' electrodes fit this criterion. These exclusion criteria left 108, 115, 92, 106, 93 and 36 electrodes for analysis for the 6 participants in Dataset 1 (Table 1) and between 76 and 228 electrodes for the 16 participants in Dataset 2 (Table 2).

Next, the common average reference (from all electrodes connected to the same amplifier) was removed for each timepoint separately. The signal in the high-gamma frequency band (70 Hz–150 Hz) was then extracted by taking the absolute value of the Hilbert transform of the signal extracted from 8 Gaussian filters (centre frequencies: 73, 79.5, 87.8, 96.9, 107, 118.1, 130.4 and 144; s.d.: 4.68, 4.92, 5.17, 5.43, 5.7, 5.99, 6.3 and 6.62, respectively, as in for example, ref. 85). The resulting envelopes from each of the Gaussian filters were averaged into one high-gamma envelope. We focused on the high-gamma frequency range because this component of the signal has been shown to track

neural activity most closely[86]. Linear interpolation was used to remove data points whose magnitude was more than 5 times the 90th percentile of all magnitudes[41], and we downsampled the signal by a factor of 4 (Matlab procedure 'resample'). For all data analyses, basic Matlab (v.2021a) functions were used.

Finally, the data were *z*-scored and normalized to a minimum/maximum value of 0/1 to allow for comparisons across electrodes, and the signal was downsampled further to 60 Hz (regardless of the participant's native sampling frequency, Matlab procedure 'resample') to reduce noise and standardize the sampling frequency across participants. For the participants who performed a slower version of the paradigm (for example, words presented for 700 ms each; see 'Experiment'), the signal was time warped to a faster rate (words presented for 450 ms each) so that timecourses could be compared across participants. This time warping was done by interpolation (Matlab procedure 'interp1').

Experiment

Dataset 1: In an event-related design, participants read sentences, lists of words, Jabberwocky sentences and lists of non-words. All stimuli were eight words/non-words long. The materials were adapted from ref. 1 and the full details of stimulus construction are described in the original publication. In short, sentences were manually constructed to cover a wide range of topics using various syntactic structures. Sentences were intended to be easily read, to fit participants with diverse clinical conditions, and only included mono- and bi-syllabic words. The full list of materials is available on OSF[37]. The word lists were created by scrambling the words from the sentences. Jabberwocky sentences were created from the sentences by removing content words (for example, nouns, verbs and so on), but leaving the syntactic frame, consisting of function words (for example, articles, conjunctions, prepositions, pronouns and so on), intact. Content words were replaced with other pronounceable non-words, matched for length (in syllables). Lastly, the non-word lists were generated by scrambling the words/non-words from the Jabberwocky condition. Originally, a set of 160 items per condition were created and here, 80 or 60 items of those were used (depending on stimulus presentation rate, as detailed below).

Each event (trial) consisted of eight words/non-words, presented one at a time at the centre 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 non-word in the Jabberwocky and Non-word-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 (Pn stands for Participant n), P5 and P6 viewed each word/non-word for 450 ms (fast timing), and P2, P3 and P4 viewed each word/non-word for 700 ms (slow timing). The presentation speed was determined before the experiment on the basis of the participant's preference. After the last word/non-word 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). Behavioural responses were continually recorded, but only responses 1 s before and 2 s after the probe were considered for calculating behavioural performance (Supplementary Table 10). Participants performed best on the Sentence trials and worst on the Non-word-list trials, with an average accuracy across all conditions of 81.01% (Supplementary Table 10). 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 and avoid predictability effects.

Trials were grouped into runs to give participants short breaks throughout the experiment. In the fast-timing version of the experiment, each run included 8 trials per condition and lasted 220 s, and in the slow-timing version, each run included 6 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. P1 was accidentally shown one run twice and consequently saw only 9 unique runs for a total of 72 trials per condition (as they opted for the fast presentation rate).

Dataset 2: In an event-related design that was similar to the one used in Dataset 1, participants read sentences and lists of non-words. The other two conditions used in Dataset 1 (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[87]. The sentences came from a language corpus (Brown corpus; ref. 88) where we searched for 12-word-long sentences and chose a diverse set among those. The non-words were created using the Wuggy software to match the words from the sentences on low-level phonology.

Each event (trial) consisted of 12 words/non-words, presented one at a time at the centre of the screen. At the end of each sequence, a memory probe was presented (a word in the Sentence condition and a non-word in the Non-word-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/non-word (medium timing) and 750 ms per word/non-word (slow timing; see Table 2 for a description of the presentation rates by participant). The presentation speed was determined before the experiment on the basis of the participant's preference. After the last word/non-word 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). Behavioural responses were continually recorded, but only responses 1 s before and 2 s after the probe were considered for calculating behavioural performance (Supplementary Table 11). As in Dataset 1, participants performed best on the Sentence trials and worse on the Non-word-list trials. However, in this sample of participants, there was substantial individual variability in the consistency and accuracy of responses (Supplementary Table 11). On average, participants provided a correct response 68.57% of the time (Supplementary Table 11). 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 slow-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. Thirteen 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/non-words of the stimulus were used to ensure that the length of the timecourses being analysed was the same across Datasets 1 and 2.

Language-responsive electrode selection

In both datasets, we identified language-responsive electrodes as electrodes that respond significantly more (on average, across trials) to sentences (the S condition) than to perceptually similar but linguistically uninformative (that is, meaningless and unstructured) non-word lists (the N condition). First, the envelope of the high-gamma signal was averaged across word/non-word positions (8 positions in the experiment used in Dataset 1 and 12 positions in the experiment used in Dataset 2) to construct an 'observed' response vector for each electrode ($1 \times n\text{TrialsS} + n\text{TrialsN}$; the number of trials, across the S and

N conditions, which 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 Non-word-list trials, a value of −1. The values in the ideal response vector were then randomly permuted without replacement and a new correlation was computed. This process was repeated 10,000 times for each electrode separately to construct a null distribution (with shuffled labels) relative to which the true correlation between the observed values and the 'idealized' values could be evaluated. Electrodes were determined to be language responsive if the observed vs idealized correlation was greater than 95% of the correlations computed using the permuted idealized response vectors (equivalent to $P < 0.05$). (We chose a liberal significance threshold to maximize the number of electrodes to be included in the critical analyses and to increase the chances of discovering distinct response profiles.) The majority of the language-responsive electrodes (98.3% in Dataset 1, 53.9% in Dataset 2) fell in the left hemisphere, but we used electrodes across both hemispheres in all analyses (see ref. 87 for evidence of a robust right-hemisphere component of the language network in a dataset of >800 participants).

Clustering analysis

Using Dataset 1 (6 participants, 177 language-responsive electrodes), we created a single timecourse per electrode by concatenating the average timecourses across the four conditions (sentences (S), word lists (W), Jabberwocky sentences (J), non-word lists (N), note that the order of the conditions concatenated did not matter since the distance metric was correlation-based). All the timepoints of the concatenated timecourses (864 data points: 60 Hz × 4 conditions × 3.6 s per trial after resampling) served as input to a *k*-medoids clustering algorithm[89]. *k*-medoids is a clustering technique that divides data points (electrodes in our case) into *k* groups, where *k* is predetermined. The algorithm attempts to minimize the distances between each electrode and the cluster centre, where cluster centres are represented by 'medoids' (exemplar electrodes selected by the algorithm) and the distance metric is correlation-based. *k*-medoids clustering was chosen over the more commonly used *k*-means clustering to allow for the use of a correlation-based distance metric as we were most interested in the shape of the timecourses over their scale which can vary due to cognitively irrelevant physiological differences (but see OSF[37] for evidence that similar clusters emerge with a *k*-means clustering algorithm using a Euclidean distance).

Optimal number of clusters

To determine the optimal number of clusters, we used the 'elbow' method[90] which searches for the value of *k* above which the increase in explained variance becomes more moderate. For each *k* (between 2 and 10), *k*-medoids clustering was performed, and explained variance was computed as the sum of the correlation-based distances of all the electrodes to their assigned cluster centre 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 above which the derivative became more moderate. We plotted the derivative of this curve as well for easier inspection of the transition point. We also repeated the elbow method while enforcing a parametrically sampled electrode-reliability threshold (from 0.1 to 0.4 in increments of 0.1) to further examine our choice of *k*. If the chosen *k* does, in fact, appropriately describe the data, we would expect the strength of the elbow (that is, the drop in explained variance for $k + 1$) to increase when more electrodes are excluded based on their lower reliability.

Partial correlation of individual electrodes with each of the cluster medoids

To evaluate the extent to which the observed responses (electrodes) can be attributed to a single profile (cluster), we computed partial correlations of every electrode's mean timecourse with that of each of the cluster medoids while controlling for the other two cluster medoids. For instance, take $r_{s1C1,C2C3}$ as the partial correlation between a signal *s*1 and Cluster 1 medoid *C*1, while controlling for the Cluster 2 medoid *C*2 and Cluster 3 medoid *C*3. $r_{s1C1,C2C3}$ can be computed by following these steps; (1) performing a multiple regression analysis with *s*1 as the dependent variable and *C*2 and *C*3 as the independent variables, obtaining the residual *e*1; (2) performing a multiple regression analysis with *C*1 as the dependent variable and *C*2 and *C*3 as the independent variable, obtaining the residual *e*2; and (3) calculating the correlation coefficient between the residuals *e*1 and *e*2. This is the partial correlation $r_{s1C1,C2C3}$. The analysis was performed using Matlab's 'partialcorr' procedure.

Cluster stability across trials

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

We then conducted a permutation analysis to statistically compare the clustering solutions. This was done separately for each of the two halves of the data (odd- and even-numbered subsets of trials). Under the null hypothesis, no distinct response profiles should be detectable in the data and consequently, responses in one electrode should be interchangeable with responses in another electrode. Using half of the data, we shuffled average responses across electrodes (within each condition separately, thus disrupting the relationship between the conditions for a given electrode while leaving the distribution of within-condition average responses intact), reclustered the electrodes into 3 clusters and then correlated the resulting cluster centres to the 'corresponding' cluster centres from the full dataset. This permutation test was determined to be more conservative than shuffling individual trials across electrodes (within each condition separately). Accordingly, comparisons remained significant when shuffling individual trials. 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.

Cluster robustness to data loss

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

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

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

Electrode locking to onsets of individual words/non-words

To estimate the degree of stimulus locking for each electrode and condition separately, we fitted a sinusoidal function that represented the stimulus train to the mean of the odd trials and then computed the Pearson correlation between the fitted sinusoidal function and the mean of the even trials. For the sinusoidal function fitting, we assumed that the frequency of the sinusoidal function was the frequency of stimulus presentation, and we fitted the phase, amplitude and offset of the sinusoid 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 an LME model, using the Matlab procedure 'fitlme', regressing the locking values of all electrodes and conditions on the fixed-effects categorical variable of 'cluster' (with 3 levels for Clusters 1, 2 or 3 according to which cluster each electrode was assigned to) and 'condition' (with 4 levels for conditions S, W, J, N), both grouped by the random-effects variable of 'participant', as well as a fixed interaction term between 'cluster' and 'condition', using the Wilkinson formula[91]:

$$\begin{aligned}&\text{Locking} \sim 1 + \text{cluster} * \text{condition}\\&+(\text{cluster}|\text{participant}) + (\text{condition}|\text{participant})\end{aligned} \quad (1)$$

An ANOVA test for LME models 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.

Electrode discrimination between conditions

To examine the timecourse of condition divergence, as quantified by the electrodes' ability to linearly discriminate between the magnitudes of pairs of conditions, we focused on condition pairs that critically differ in their engagement of particular linguistic processes: conditions S and W, which differ in whether they engage combinatorial (syntactic and semantic) processing (S=yes, W=no), conditions W and N, which differ in whether they engage word meaning processing (W=yes, N=no), and conditions J and N, which differ in whether they engage syntactic processing (J=yes, N=no). This analysis tests how early the relevant pair of conditions reliably diverge and the strength of that divergence. For every electrode, the mean response to the three conditions of interest (S, W and N) was averaged across 100 ms bins with a 100 ms sliding window. For each cluster separately, a binary logistic classifier (selected from the best of 20 random instantiations; performed using Matlab's 'fitclinear' procedure) was trained to discriminate S from W, W from N, or J 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 accuracy (the percent of mean responses from all electrodes in a particular cluster correctly classified as S/W, W/N, or J/N). The performance of the model at a given time bin was statistically evaluated using a cluster permutation test to control for multiple comparisons and account for the autocorrelation structure of the signals[38,39]. This was done by shuffling the condition labels 1,000 times for each time bin to simulate surrogate data. For each surrogate data repetition, we computed the sum of consecutive t-values that passed some arbitrary t-value threshold, referred to as the t-sum statistics. We chose a t-value threshold corresponding to an alpha level of 0.05. Using the t-sum values from the 1,000 permutations, we constructed a null distribution for this t-sum statistic and then compared it to the same t-sum statistic computed from the real data to assess significance.

Computing *n*-gram frequencies of Sentence and Non-word stimuli

N-gram frequencies were extracted from the Google *n*-gram online platform (https://books.google.com/ngrams/), averaging across Google books corpora between the years 2010 and 2020. The *n*-gram frequency for $n = 1$ is the frequency of that individual word in the corpus; the *n*-gram frequency for $n = 2$ is the frequency of the bigram (sequence of 2 words) ending in, and including, that word; the *n*-gram frequency for $n = 3$ is the frequency of the trigram (sequence of 3 words) ending in, and including, that word and so on. Sequences that were not found in the corpus were assigned a value of 0.

Estimation of temporal receptive window size per electrode

We used a simplified model to simulate neural responses in the Sentence (S) condition as a convolution of a stimulus train and truncated Gaussian kernels with varying widths. The kernels represented an evoked 'response function' with a width (σ) corresponding to the temporal receptive window (TRW) of an idealized neural population underlying the intracranial responses measured by a single electrode. The kernels were constructed from Gaussian curves with a standard deviation of $\sigma/2$ truncated at ±1 s.d. (capturing 2/3 of the area under the Gaussian). We then normalized the truncated Gaussians to have a minimum of 0 and maximum of 1. We chose a symmetric kernel because we wanted to capture the full assumed TRW for a straightforward interpretation of the fitted window size. For instance, a long-tailed response function would have a shorter 'effective' receptive window because the tails of the kernel would affect the neural response much less than the centre of the kernel. We further chose a kernel with smooth edges because we assumed that neural activity in response to a stimulus would increase and decrease gradually (cf. an abrupt change of voltage such as in a boxcar shape), given that macroelectrodes sum activity from a large neural population[86]. Furthermore, we assumed that the TRW for a given neural population was 'fixed', but see Discussion.

We also verified that the specific shape of kernel did not affect our main result. We tested five different response functions: cosine, 'wide' Gaussian (Gaussian curves with a standard deviation of $\sigma/2$ that were truncated at ±1 s.d., as used in the manuscript), 'narrow' Gaussian (Gaussian curves with a standard deviation of $\sigma/16$ that were truncated at ±8 s.d.), a square (that is, boxcar) function (1 for the entire window) and a linear asymmetric function (linear function with a value of 0 initially and a value of 1 at the end of the window).

The stimulus train took a value of 1 at the time of word onsets and 0 otherwise, assuming, for simplicity, that the minimal stimulus unit

of interest for language-responsive neural populations is a word (cf. for example, refs. 53,25 for evidence that the language network is sensitive to sublexical structure). Neural responses were simulated for σ ranging from one-third of a word to 8 words (the length of our stimuli), in 1 sample increments (1/27th of a word, the highest resolution we were able to evaluate given our sampling rate of 60 Hz). Our implementation of the convolution is identical to assuming that the kernels appear as evoked responses starting at each word onset (see OSF[37]). The length of the evoked response/kernel is directly mapped onto a longer temporal receptive window, such that when a stimulus evokes a wider response, its effect will remain for a longer period of time.

To find the best fit of the temporal receptive window size for each electrode after simulating neural signals using this toy model, we selected the TRW size that yielded the highest correlation between the simulated neural response (also normalized to be between 0 and 1) and the actual neural response. The value of the correlation was taken as a proxy for the goodness of fit.

To evaluate significance, we ran LME models regressing the estimates of temporal receptive window sizes (σ) of all electrodes on the fixed-effects categorical variable of 'cluster' grouped by the random-effects variable of 'participant'. Cluster was dummy coded as a categorical variable with three levels, and Cluster 1 was treated as the baseline intercept. This approach allowed us to compare electrodes in Cluster 2 to those in Cluster 1, and electrodes in Cluster 3 to those in Cluster 1. To additionally compare electrodes in Clusters 2 vs 3, we ran another similar LME model with the only difference being that the baseline intercept was now the Cluster 2 category (Supplementary Tables 5–9). To account for the small number of participants in Dataset 1, we used the Satterthwaite method[92].

Anatomical topography analysis

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

To do this, we extracted the MNI coordinates of all the electrodes in each of the three clusters and ran LME models regressing each of the coordinates (either y or z) on the fixed-effects categorical variable of 'cluster' grouped by the random-effects variable of 'participant', using the Wilkinson formula[91]:

$$\text{Coordinate} \sim 1 + \text{cluster} + (1 + \text{cluster}|\text{participant}) \quad (2)$$

where 'Coordinate' is either the y or z MNI coordinate. The random effect that groups the estimates by participant ensures that electrode coordinates are compared within participants. This approach is crucial for accommodating inter-individual variability in the precise locations of language areas (for example, ref. 1), which means that the absolute values of MNI coordinates cannot be easily compared between participants.

Cluster was dummy coded as a categorical variable with three levels and Cluster 1 was treated as the baseline intercept. This approach allowed us to compare electrodes in Cluster 2 to those in Cluster 1 and electrodes in Cluster 3 to those in Cluster 1. To additionally compare electrodes in Clusters 2 vs 3, we ran another similar LME model with the only difference being that the baseline intercept was now the Cluster 2 category (Supplementary Tables 2–4). To account for the small number of participants in Dataset 1, we used the Satterthwaite corrective degree-of-freedom approximation method, combined with REML fitting for LME, which was shown to be most effective when using the Satterthwaite method[92].

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

Replication of the clusters in Dataset 2

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

We then applied the same clustering procedure to Dataset 2 (16 participants, 362 language-responsive electrodes). The elbow method revealed that the optimal number of clusters in Dataset 2 is $k = 2$. However, because the optimal number of clusters in Dataset 1 was $k = 3$, we examined the clustering solutions for both $k = 2$ and $k = 3$. We also performed an analysis where we assigned electrodes in Dataset 2 to the most correlated Dataset 1 cluster (also termed "winner-take-all"). This analysis was intended to examine whether responses such as those found in Dataset 1 were at all present in Dataset 2 (even if they did not emerge as strongly through clustering); thus, the assignment of electrodes to a 'group' was done by correlation alone, and no actual clustering was performed.

To statistically compare the clustering solutions between Datasets 1 and 2 for $k = 3$, we used the same approach as the one described above ('Stability of clusters across trials'). In particular, using Dataset 2, we shuffled average responses across electrodes (within each condition separately), reclustered the electrodes into 3 clusters and then correlated the resulting cluster averages to the cluster averages 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 centres from the permuted data to the two clusters in Dataset 1 that were most strongly correlated with the two clusters that emerged from Dataset 2 (that is, Clusters 1 and 3).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Preprocessed data, all stimuli and statistical results, as well as selected additional analyses are available on OSF at https://osf.io/xfbr8/ (ref. 37). Raw data may be provided upon request to the corresponding authors and institutional approval of a data-sharing agreement.

Code availability

Code used to conduct analyses and generate figures from the preprocessed data is available publicly on GitHub at https://github.com/coltoncasto/ecog_clustering_PUBLIC (ref. 93). The VERA software suite used to perform electrode localization can also be found on GitHub at https://github.com/neurotechcenter/VERA (ref. 82).

References

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

2. Pallier, C., Devauchelle, A. D. & Dehaene, S. Cortical representation of the constituent structure of sentences. *Proc. Natl Acad. Sci. USA* **108**, 2522–2527 (2011).
3. Regev, M., Honey, C. J., Simony, E. & Hasson, U. Selective and invariant neural responses to spoken and written narratives. *J. Neurosci.* **33**, 15978–15988 (2013).
4. Scott, T. L., Gallée, J. & Fedorenko, E. A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cogn. Neurosci.* **8**, 167–176 (2017).
5. Diachek, E., Blank, I., Siegelman, M., Affourtit, J. & Fedorenko, E. 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 (2020).
6. Malik-Moraleda, S. et al. An investigation across 45 languages and 12 language families reveals a universal language network. *Nat. Neurosci.* **25**, 1014–1019 (2022).
7. Fedorenko, E., Behr, M. K. & Kanwisher, N. Functional specificity for high-level linguistic processing in the human brain. *Proc. Natl Acad. Sci. USA* **108**, 16428–16433 (2011).
8. Monti, M. M., Parsons, L. M. & Osherson, D. N. Thought beyond language: neural dissociation of algebra and natural language. *Psychol. Sci.* **23**, 914–922 (2012).
9. Deen, B., Koldewyn, K., Kanwisher, N. & Saxe, R. Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb. Cortex* **25**, 4596–4609 (2015).
10. Ivanova, A. A. et al. The language network is recruited but not required for nonverbal event semantics. *Neurobiol. Lang.* **2**, 176–201 (2021).
11. Chen, X. et al. The human language system, including its inferior frontal component in "Broca's area," does not support music perception. *Cereb. Cortex* **33**, 7904–7929 (2023).
12. Fedorenko, E., Ivanova, A. A. & Regev, T. I. The language network as a natural kind within the broader landscape of the human brain. *Nat. Rev. Neurosci.* **25**, 289–312 (2024).
13. Okada, K. & Hickok, G. Identification of lexical-phonological networks in the superior temporal sulcus using functional magnetic resonance imaging. *Neuroreport* **17**, 1293–1296 (2006).
14. Graves, W. W., Grabowski, T. J., Mehta, S. & Gupta, P. The left posterior superior temporal gyrus participates specifically in accessing lexical phonology. *J. Cogn. Neurosci.* **20**, 1698–1710 (2008).
15. DeWitt, I. & Rauschecker, J. P. Phoneme and word recognition in the auditory ventral stream. *Proc. Natl Acad. Sci. USA* **109**, E505–E514 (2012).
16. Price, C. J., Moore, C. J., Humphreys, G. W. & Wise, R. J. S. Segregating semantic from phonological processes during reading. *J. Cogn. Neurosci.* **9**, 727–733 (1997).
17. Mesulam, M. M. et al. Words and objects at the tip of the left temporal lobe in primary progressive aphasia. *Brain* **136**, 601–618 (2013).
18. Friederici, A. D. The brain basis of language processing: from structure to function. *Physiol. Rev.* **91**, 1357–1392 (2011).
19. Hagoort, P. On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* **9**, 416–423 (2005).
20. Grodzinsky, Y. & Santi, A. The battle for Broca's region. *Trends Cogn. Sci.* **12**, 474–480 (2008).
21. Matchin, W. & Hickok, G. The cortical organization of syntax. *Cereb. Cortex* **30**, 1481–1498 (2020).
22. Fedorenko, E., Blank, I. A., Siegelman, M. & Mineroff, Z. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition* **203**, 104348 (2020).
23. Bautista, A. & Wilson, S. M. Neural responses to grammatically and lexically degraded speech. *Lang. Cogn. Neurosci.* **31**, 567–574 (2016).
24. Anderson, A. J. et al. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *J. Neurosci.* **41**, 4100–4119 (2021).
25. Regev, T. I. et al. High-level language brain regions process sublexical regularities. *Cereb. Cortex* **34**, bhae077 (2024).
26. Mukamel, R. & Fried, I. Human intracranial recordings and cognitive neuroscience. *Annu. Rev. Psychol.* **63**, 511–537 (2011).
27. Fedorenko, E. et al. Neural correlate of the construction of sentence meaning. *Proc. Natl Acad. Sci. USA* **113**, E6256–E6262 (2016).
28. Nelson, M. J. et al. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proc. Natl Acad. Sci. USA* **114**, E3669–E3678 (2017).
29. Woolnough, O. et al. Spatiotemporally distributed frontotemporal networks for sentence reading. *Proc. Natl Acad. Sci. USA* **120**, e2300252120 (2023).
30. Desbordes, T. et al. Dimensionality and ramping: signatures of sentence integration in the dynamics of brains and deep language models. *J. Neurosci.* **43**, 5350–5364 (2023).
31. Goldstein, A. et al. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
32. Lerner, Y., Honey, C. J., Silbert, L. J. & Hasson, U. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).
33. Blank, I. A. & Fedorenko, E. No evidence for differences among language regions in their temporal receptive windows. *Neuroimage* **219**, 116925 (2020).
34. Jain, S. et al. Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. In *NeurIPS Proc. Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (eds Larochelle, H. et al.) 1–12 (NeurIPS, 2020).
35. Fedorenko, E., Nieto-Castañon, A. & Kanwisher, N. Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia* **50**, 499–513 (2012).
36. Shain, C. et al. Distributed sensitivity to syntax and semantics throughout the human language network. *J. Cogn. Neurosci.* **36**, 1427–1471 (2024).
37. Regev, T. I., Casto, C. & Fedorenko, E. Neural populations in the language network differ in the size of their temporal receptive windows. *OSF* osf.io/xfbr8 (2024).
38. Stelzer, J., Chen, Y. & Turner, R. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *Neuroimage* **65**, 69–82 (2013).
39. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
40. Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550 (2008).
41. Norman-Haignere, S. V. et al. Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nat. Hum. Behav.* **6**, 455–469 (2022).
42. Overath, T., McDermott, J. H., Zarate, J. M. & Poeppel, D. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat. Neurosci.* **18**, 903–911 (2015).
43. Keshishian, M. et al. Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nat. Hum. Behav.* **7**, 740–753 (2023).
44. Braga, R. M., DiNicola, L. M., Becker, H. C. & Buckner, R. L. Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *J. Neurophysiol.* **124**, 1415–1448 (2020).

45. Fedorenko, E. & Blank, I. A. Broca's area is not a natural kind. *Trends Cogn. Sci.* **24**, 270–284 (2020).
46. Dick, F. et al. 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 (2001).
47. Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).
48. Murray, J. D. et al. A hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.* **17**, 1661–1663 (2014).
49. Chien, H. S. & Honey, C. J. Constructing and forgetting temporal context in the human cerebral cortex. *Neuron* **106**, 675–686 (2020).
50. Jacoby, N. & Fedorenko, E. Discourse-level comprehension engages medial frontal Theory of Mind brain regions even for expository texts. *Lang. Cogn. Neurosci.* **35**, 780–796 (2018).
51. Caucheteux, C., Gramfort, A. & King, J. R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* **7**, 430–441 (2023).
52. Chang, C. H. C., Nastase, S. A. & Hasson, U. Information flow across the cortical timescale hierarchy during narrative construction. *Proc. Natl Acad. Sci. USA* **119**, e2209307119 (2022).
53. Bozic, M., Tyler, L. K., Ives, D. T., Randall, B. & Marslen-Wilson, W. D. Bihemispheric foundations for human speech comprehension. *Proc. Natl Acad. Sci. USA* **107**, 17439–17444 (2010).
54. Paulk, A. C. et al. Large-scale neural recordings with single neuron resolution using Neuropixels probes in human cortex. *Nat. Neurosci.* **25**, 252–263 (2022).
55. Leonard, M. K. et al. Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature* **626**, 593–602 (2024).
56. Evans, N. & Levinson, S. C. The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.* **32**, 429–448 (2009).
57. Shannon, C. E. Communication in the presence of noise. *Proc. IRE* **37**, 10–21 (1949).
58. Levy, R. Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).
59. Levy, R. A noisy-channel model of human sentence comprehension under uncertain input. In *Proc. 2008 Conference on Empirical Methods in Natural Language Processing* (eds Lapata, M. & Ng, H. T.) 234–243 (Association for Computational Linguistics, 2008).
60. Gibson, E., Bergen, L. & Piantadosi, S. T. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proc. Natl Acad. Sci. USA* **110**, 8051–8056 (2013).
61. Keshev, M. & Meltzer-Asscher, A. Noisy is better than rare: comprehenders compromise subject–verb agreement to form more probable linguistic structures. *Cogn. Psychol.* **124**, 101359 (2021).
62. Gibson, E. et al. How efficiency shapes human language. *Trends Cogn. Sci.* **23**, 389–407 (2019).
63. Tuckute, G., Kanwisher, N. & Fedorenko, E. Language in brains, minds, and machines. *Annu. Rev. Neurosci.* https://doi.org/10.1146/annurev-neuro-120623-101142 (2024).
64. Norman-Haignere, S., Kanwisher, N. G. & McDermott, J. H. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* **88**, 1281–1296 (2015).
65. Baker, C. I. et al. Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proc. Natl Acad. Sci. USA* **104**, 9087–9092 (2007).
66. Buckner, R. L. & DiNicola, L. M. The brain's default network: updated anatomy, physiology and evolving insights. *Nat. Rev. Neurosci.* **20**, 593–608 (2019).
67. Saxe, R., Brett, M. & Kanwisher, N. Divide and conquer: a defense of functional localizers. *Neuroimage* **30**, 1088–1096 (2006).
68. Baldassano, C. et al. Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721 (2017).
69. Wilson, S. M. et al. Recovery from aphasia in the first year after stroke. *Brain* **146**, 1021–1039 (2023).
70. Piantadosi, S. T., Tily, H. & Gibson, E. Word lengths are optimized for efficient communication. *Proc. Natl Acad. Sci. USA* **108**, 3526–3529 (2011).
71. Shain, C., Blank, I. A., Fedorenko, E., Gibson, E. & Schuler, W. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *J. Neurosci.* **42**, 7412–7430 (2022).
72. Schrimpf, M. et al. The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl Acad. Sci. USA* **118**, e2105646118 (2021).
73. Tuckute, G. et al. Driving and suppressing the human language network using large language models. *Nat. Hum. Behav.* **8**, 544–561 (2024).
74. Mollica, F. & Piantadosi, S. T. Humans store about 1.5 megabytes of information during language acquisition. *R. Soc. Open Sci.* **6**, 181393 (2019).
75. Skrill, D. & Norman-Haignere, S. V. Large language models transition from integrating across position-yoked, exponential windows to structure-yoked, power-law windows. *Adv. Neural Inf. Process. Syst.* **36**, 638–654 (2023).
76. Giglio, L., Ostarek, M., Weber, K. & Hagoort, P. Commonalities and asymmetries in the neurobiological infrastructure for language production and comprehension. *Cereb. Cortex* **32**, 1405–1418 (2022).
77. Hu, J. et al. Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cereb. Cortex* **33**, 4384–4404 (2023).
78. Lee, E. K., Brown-Schmidt, S. & Watson, D. G. Ways of looking ahead: hierarchical planning in language production. *Cognition* **129**, 544–562 (2013).
79. Wechsler, D. Wechsler abbreviated scale of intelligence (WASI) [Database record]. *APA PsycTests* https://psycnet.apa.org/doi/10.1037/t15170-000 (APA PsycNet, 1999).
80. Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N. & Wolpaw, J. R. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Trans. Biomed. Eng.* **51**, 1034–1043 (2004).
81. Adamek, M., Swift, J. R. & Brunner, P. VERA - Versatile Electrode Localization Framework. *Zenodo* https://doi.org/10.5281/zenodo.7486842 (2022).
82. Adamek, M., Swift, J. R. & Brunner, P. VERA - A Versatile Electrode Localization Framework (Version 1.0.0). *GitHub* https://github.com/neurotechcenter/VERA (2022).
83. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).
84. Janca, R. et al. Detection of interictal epileptiform discharges using signal envelope distribution modelling: application to epileptic and non-epileptic intracranial recordings. *Brain Topogr.* **28**, 172–183 (2015).
85. Dichter, B. K., Breshears, J. D., Leonard, M. K. & Chang, E. F. The control of vocal pitch in human laryngeal motor cortex. *Cell* **174**, 21–31 (2018).
86. Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J. & Hsiao, S. S. Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *J. Neurosci.* **28**, 11526–11536 (2008).

87. Lipkin, B. et al. Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Sci. Data* **9**, 529 (2022).
88. Kučera, H. *Computational Analysis of Present-day American English* (Univ. Pr. of New England, 1967).
89. Kaufman, L. & Rousseeuw, P. J. in *Finding Groups in Data: An Introduction to Cluster Analysis* (eds L. Kaufman, L. & Rousseeuw, P. J.) Ch. 2 (Wiley, 1990).
90. Rokach, L. & Maimon, O. in *The Data Mining and Knowledge Discovery Handbook* (eds Maimon, O. & Rokach, L.) 321–352 (Springer, 2005).
91. Wilkinson, G.N. & Rogers, C.E. Symbolic description of factorial models for analysis of variance. *J. R. Stat. Soc., C: Appl.Stat.* **22**, 392–399 (1973).
92. Luke, S. G. Evaluating significance in linear mixed-effects models in R. *Behav. Res. Methods* **49**, 1494–1502 (2017).
93. Regev, T. I. et al. Neural populations in the language network differ in the size of their temporal receptive windows. *GitHub* https://github.com/coltoncasto/ecog_clustering_PUBLIC (2024).

Acknowledgements

We thank the participants for agreeing to take part in our study, as well as N. Kanwisher, former and current EvLab members, especially C. Shain and A. Ivanova, and the audience at the Neurobiology of Language conference (2022, Philadelphia) for helpful discussions and comments on the analyses and manuscript. T.I.R. was supported by the Zuckerman-CHE STEM Leadership Program and by the Poitras Center for Psychiatric Disorders Research. C.C. was supported by the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. A.L.R. was supported by NIH award U01-NS108916. J.T.W. was supported by NIH awards R01-MH120194 and P41-EB018783, and the American Epilepsy Society Research and Training Fellowship for clinicians. P.B. was supported by NIH awards R01-EB026439, U24-NS109103, U01-NS108916, U01-NS128612 and P41-EB018783, the McDonnell Center for Systems Neuroscience, and Fondazione Neurone. E.F. was supported by NIH awards R01-DC016607, R01-DC016950 and U01-NS121471, and research funds from the McGovern Institute for Brain Research, Brain and Cognitive Sciences Department, and the Simons Center for the Social Brain. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

T.I.R. and C.C. equally contributed to study conception and design, data analysis and interpretation of results, and manuscript writing. E.A.H. contributed to data analysis and manuscript editing; M.A. to data collection and analysis; A.L.R., J.T.W. and P.B. to data collection and manuscript editing. E.F. contributed to study conception and design, supervision, interpretation of results and manuscript writing.

Competing interests

The authors declare no competing interests.

Additional information

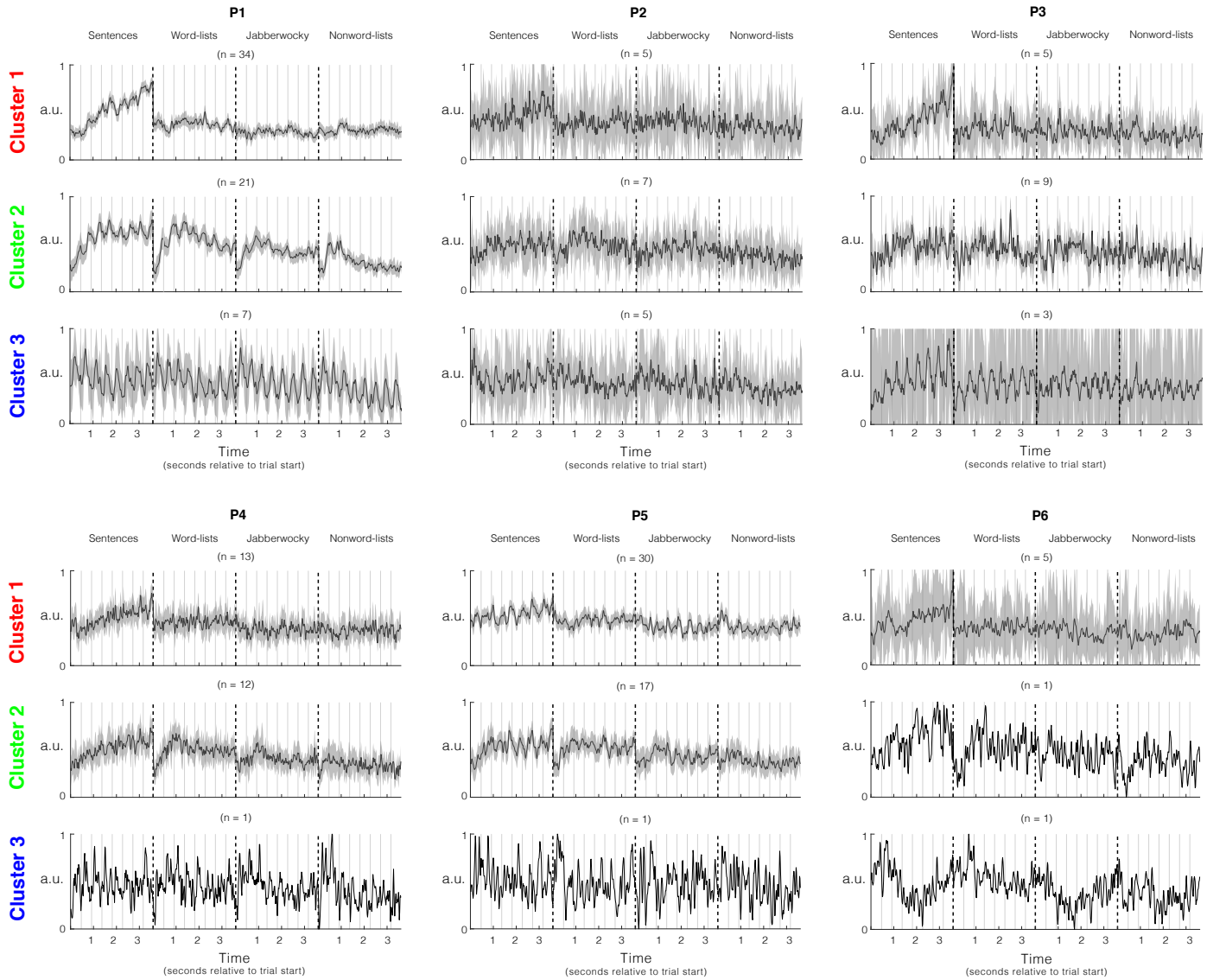
Extended data is available for this paper at https://doi.org/10.1038/s41562-024-01944-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-01944-2.

Correspondence and requests for materials should be addressed to Tamar I. Regev, Colton Casto or Evelina Fedorenko.

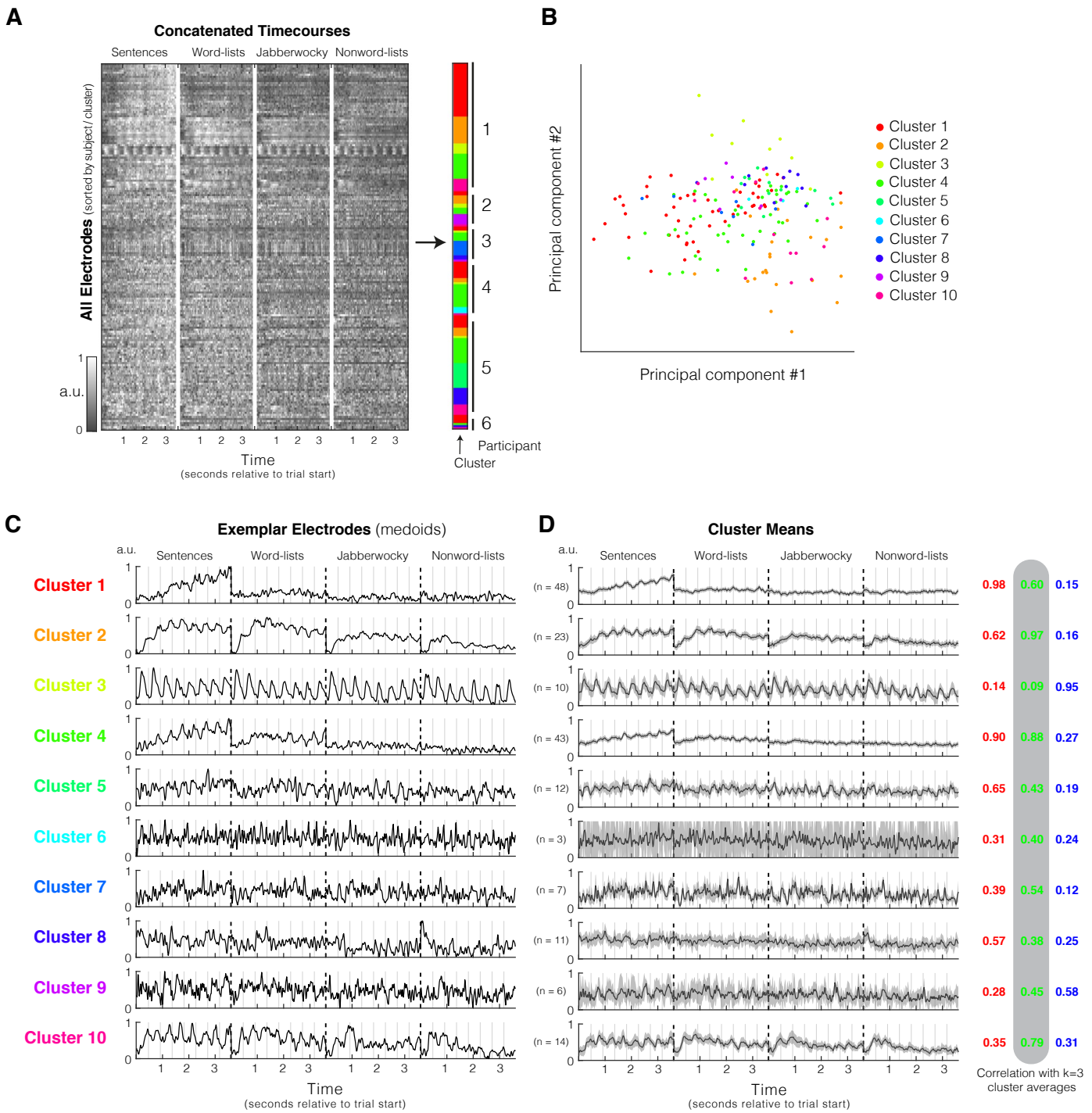
Peer review information *Nature Human Behaviour* thanks Nima Mesgarani, Jonathan Venezia and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41562-024-01944-2



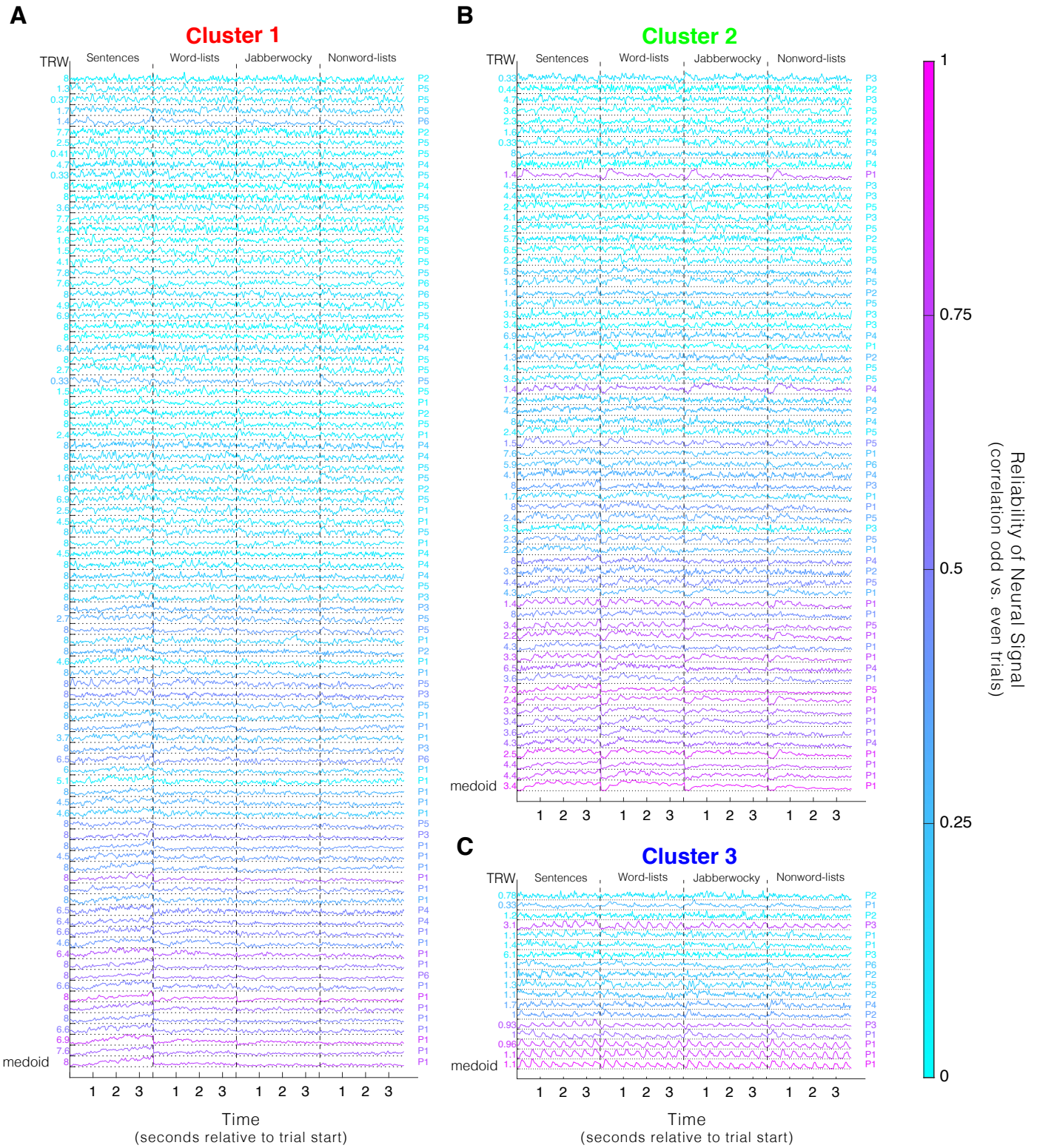
Extended Data Fig. 1 | Dataset 1 k-medoids (k = 3) cluster assignments by participant. Average cluster responses as in Fig. 2e grouped by participant. Shaded areas around the signal reflect a 99% confidence interval over electrodes. The number of electrodes constructing the average (n) is denoted above each signal in parenthesis. Prototypical responses for each of the three clusters were found in nearly all participants individually. However, for participants with only a few electrodes assigned to a given cluster (for example, P5 Cluster 3), the responses were more variable.

https://doi.org/10.1038/s41562-024-01944-2



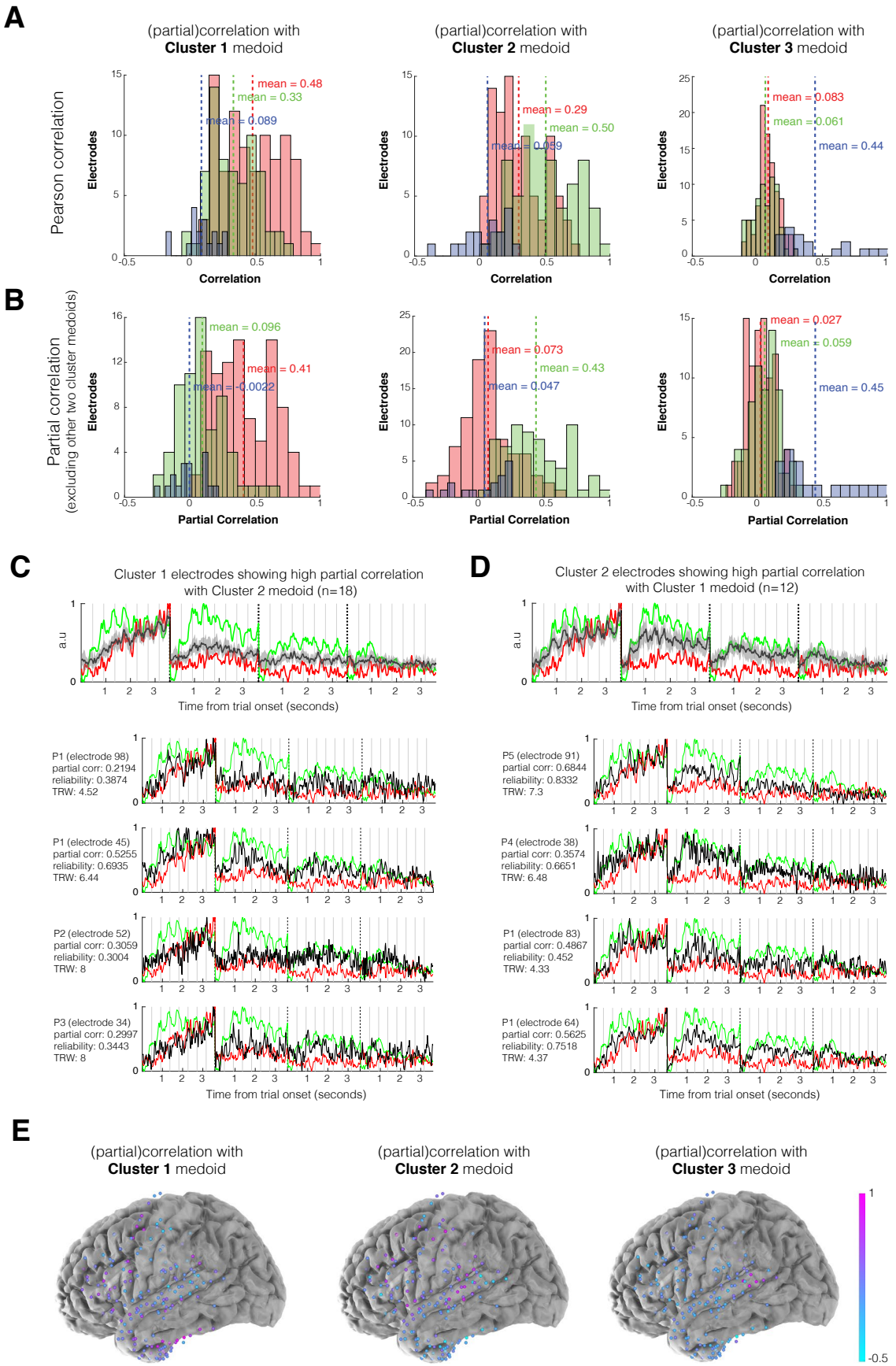
Extended Data Fig. 2 | Dataset 1 k-medoids clustering with k = 10. a) Clustering mean electrode responses (S + W + J + N) using k-medoids with k = 10 and a correlation-based distance. Shading of the data matrix reflects normalized high-gamma power (70–150 Hz). **b)** Electrode responses visualized on their first two principal components, colored by cluster. **c)** Timecourses of best representative electrodes ('medoids') selected by the algorithm from each of the ten clusters. **d)** Timecourses averaged across all electrodes in each cluster. Shaded areas around the signal reflect a 99% confidence interval over electrodes. Correlation with the k = 3 cluster averages are shown to the right of the timecourses. Many clusters exhibited high correlations with the k = 3 response profiles from Fig. 2.

https://doi.org/10.1038/s41562-024-01944-2



Extended Data Fig. 3 | All Dataset 1 responses. a-c) All Dataset 1 electrode responses. The timecourses (concatenated across the four conditions, ordered: sentences, word lists, Jabberwocky sentences, non-word lists) of all electrodes in Dataset 1 sorted by their correlation to the cluster medoid (medoid shown at the bottom of each cluster). Colors reflect the reliability of the measured neural signal, computed by correlating responses to odd and even trials (Fig. 1d). The estimated temporal receptive window (TRW) using the toy model from Fig. 4 is displayed to the left, and the participant who contributed the electrode is displayed to the right. There was strong consistency in the responses from individual electrodes within a cluster (with more variability in the less reliable electrodes), and electrodes with responses that were more similar to the cluster medoid tended to be more reliable (more pink). Note that there were two reliable response profiles (relatively pink) that showed a pattern that was distinct from the three prototypical response profiles: One electrode in Cluster 2 (the 10th electrode from the top in panel B) responded only to the onset of the first word/nonword in each trial; and one electrode in Cluster 3 (the 4th electrode from the top in panel C) was highly locked to all onsets except the first word/nonword. These profiles indicate that although the prototypical clusters explain a substantial amount of the functional heterogeneity of responses in the language network, they were not the only observed responses.

https://doi.org/10.1038/s41562-024-01944-2

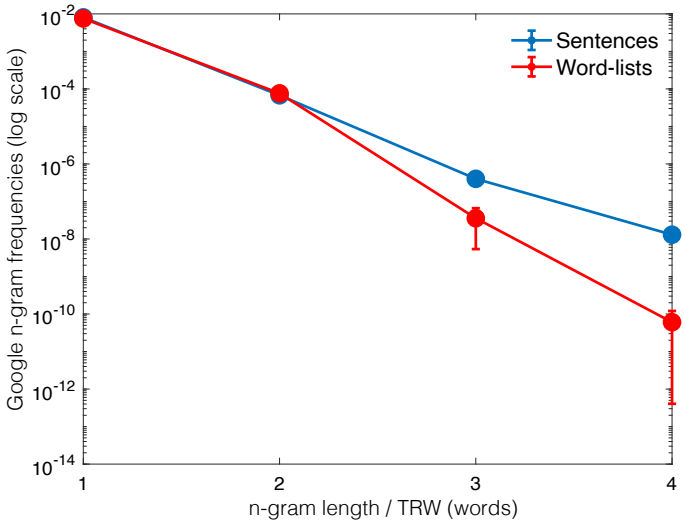


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41562-024-01944-2

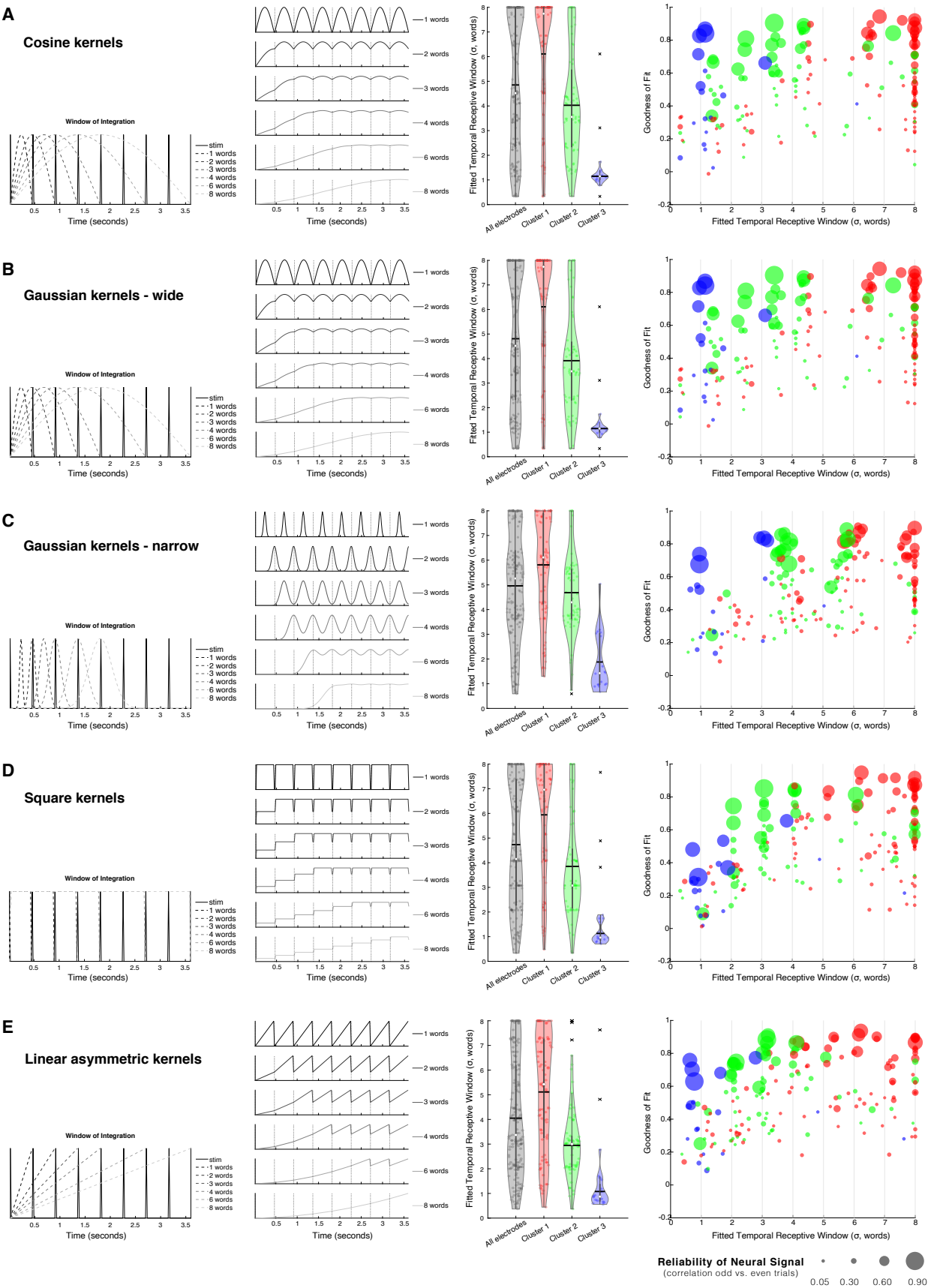
Extended Data Fig. 4 | Partial correlations of individual response profiles with each of the cluster medoids. a) Pearson correlations of all response profiles with each of the cluster medoids, grouped by cluster assignment. **b)** Partial correlations (Methods) of all response profiles with each of the cluster medoids, controlling for the other two cluster medoids, grouped by cluster assignment. **c)** Response profiles from electrodes assigned to Cluster 1 that had a high partial correlation (> 0.2, arbitrarily defined threshold) with the Cluster 2 medoid (and split-half reliability>0.3). ***Top:*** Average over all electrodes that met these criteria (n = 18, black). The Cluster 1 medoid is shown in red, and the Cluster 2 medoid is shown in green. ***Bottom:*** Four sample electrodes (black). **d)** Response profiles assigned to Cluster 2 that had a high partial correlation (> 0.2, arbitrarily defined threshold) with the Cluster 1 medoid (and split-half reliability>0.3). ***Top:*** Average over all electrodes that meet these criteria (n = 12, black). The Cluster 1 medoid is shown in red, and the Cluster 2 medoid is shown in green. ***Bottom:*** Four sample electrodes (black; see osf.io/xfbr8/ for all mixed response profiles with split-half reliability>0.3). **e)** Anatomical distribution of electrodes in Dataset 1 colored by their partial correlation with a given cluster medoid (controlling for the other two medoids). Cluster-1- and Cluster-2-like responses were present throughout frontal and temporal areas (with Cluster 1 responses having a slightly higher concentration in the temporal pole and Cluster 2 responses having a slightly higher concentration in the superior temporal gyrus (STG)), whereas Cluster-3-like responses were localized to the posterior STG.

https://doi.org/10.1038/s41562-024-01944-2



Extended Data Fig. 5 | N-gram frequencies of sentences and word lists diverge with n-gram length. N-gram frequencies were extracted from the Google n-gram online platform (https://books.google.com/ngrams/), averaging across Google books corpora between the years 2010 and 2020. For each individual word, the n-gram frequency for n = 1 was the frequency of that word in the corpus; for n = 2 it was the frequency of the bigram (sequence of 2 words) ending in that word; for n = 3 it was the frequency of the trigram (sequence of 3 words) ending in that word; and so on. Sequences that were not found in the corpus were assigned a value of 0. Results are only presented until n = 4 because for n > 4 most of the string sequences, both from the Sentence and Word-list conditions, were not found in the corpora. The plot shows that the difference between the log n-gram values for the sentences and word lists in our stimulus set grows as a function of N. Error bars represent the standard error of the mean across all n-grams extracted from the stimuli used (640, 560, 480, 399 n-grams for n-gram length = 1, 2, 3, and 4, respectively).

https://doi.org/10.1038/s41562-024-01944-2

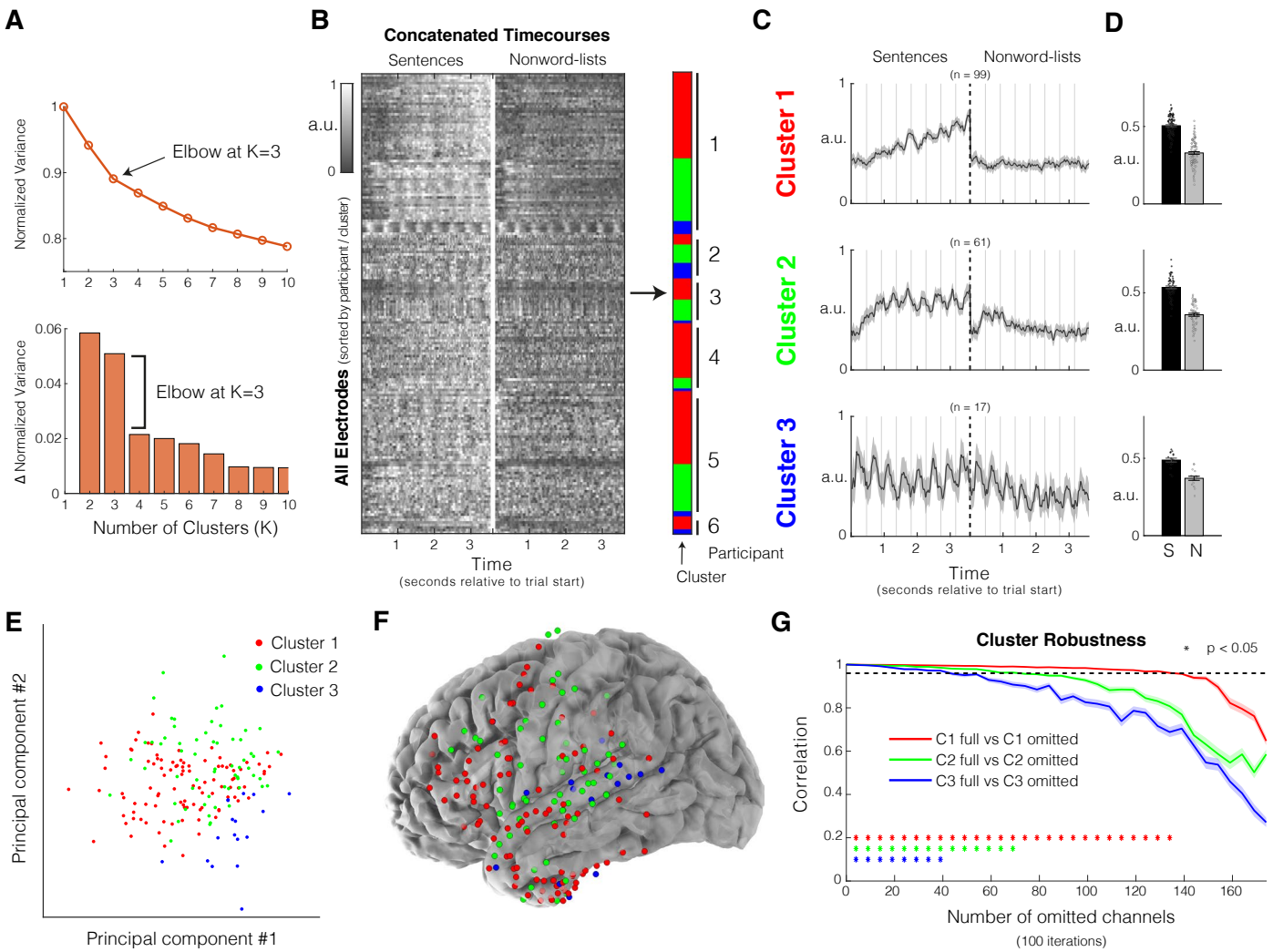


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41562-024-01944-2

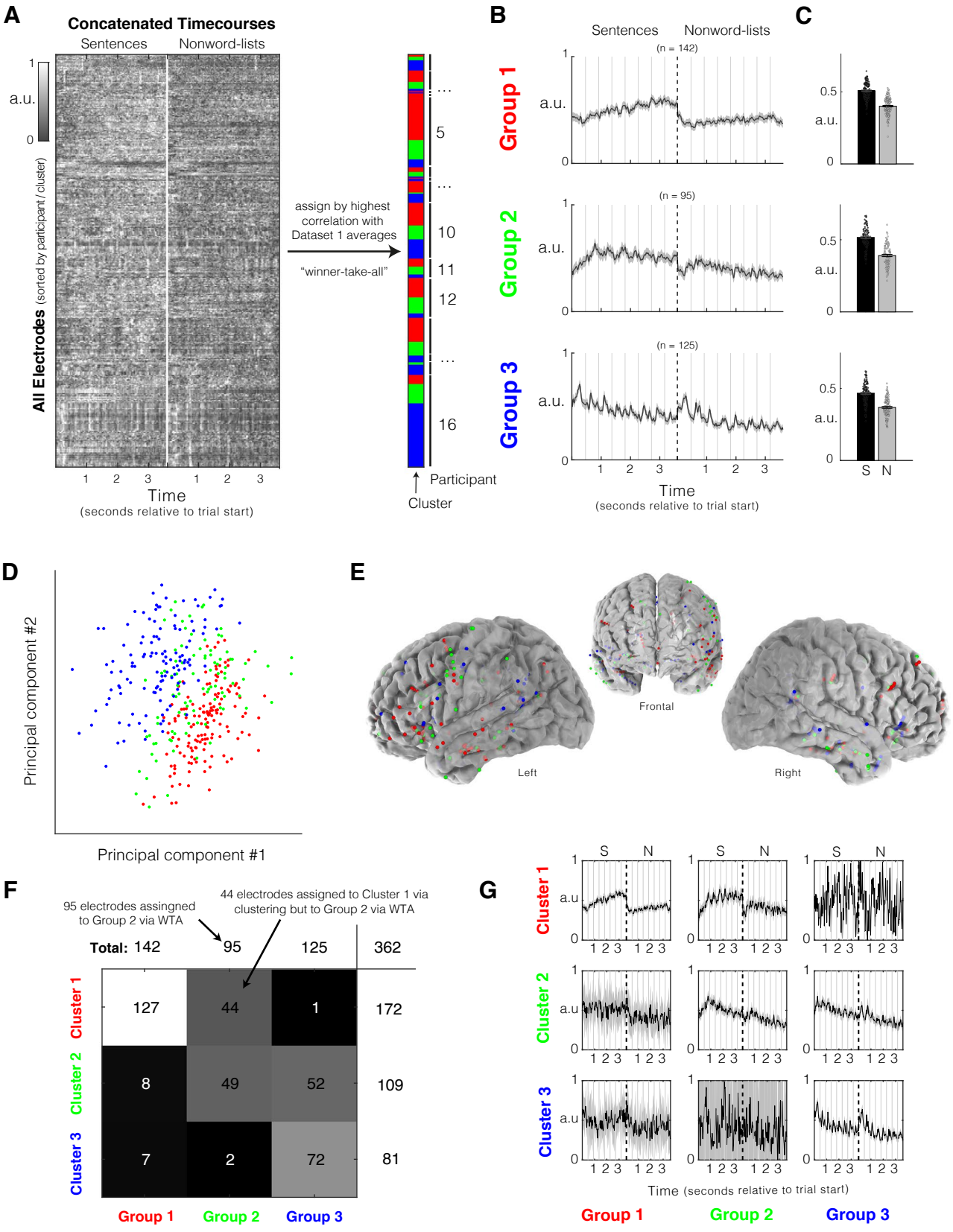
Extended Data Fig. 6 | Temporal receptive window (TRW) estimates with kernels of different shapes. The toy TRW model from Fig. 4 was applied using five different kernel shapes: cosine (**a**), 'wide' Gaussian (Gaussian curves with a standard deviation of $\sigma/2$ that were truncated at +/−1 standard deviation, as used in Fig. 4; **b**), 'narrow' Gaussian (Gaussian curves with a standard deviation of $\sigma/16$ that were truncated at +/− 8 standard deviations; **c**), a square (that is, boxcar) function (1 for the entire window; **d**) and a linear asymmetric function (linear function with a value of 0 initially and a value of 1 at the end of the window; **e**). For each kernel (**a**-**e**), the plots represent (left to right, all details are identical to Fig. 4 in the manuscript): 1) The kernel shapes for TRW = 1, 2, 3, 4, 6 and 8 words, superimposed on the simplified stimulus train; 2) The simulated neural signals for each of those TRWs; 3) violin plots of best fitted TRW values across electrodes (each dot represents an electrode, horizontal black lines are means across the electrodes, white dots are medians, vertical thin box represents lower and upper quartiles and 'x' marks indicate outliers; more than 1.5 interquartile ranges above the upper quartile or less than 1.5 interquartile ranges below the lower quartile) for all electrodes (black), or electrodes from only Clusters 1 (red) 2 (green) or 3 (blue); and 4) Estimated TRW as a function of goodness of fit. Each dot is an electrode, its size represents the reliability of its neural response, computed via correlation between the mean signals when using only odd vs. only even trials, x-axis is the electrode's best fitted TRW, y-axis is the goodness of fit, computed via correlation between the neural signal and the closest simulated signal. For all kernels the TRWs showed a decreasing trend from Cluster 1 to 3.

https://doi.org/10.1038/s41562-024-01944-2



Extended Data Fig. 7 | Dataset 1 k-medoids clustering results with only S and N conditions. a) Search for optimal k using the 'elbow method'. ***Top:*** variance (sum of the distances of all electrodes to their assigned cluster centre) normalized by the variance when k = 1 as a function of k (normalized variance (NV)). ***Bottom:*** change in NV as a function of k (NV(k + 1) − NV(k)). After k = 3 the change in variance became more moderate, suggesting that 3 clusters appropriately described Dataset 1 when using only the responses to sentences and non-words (as was the case when all four conditions were used). **b)** Clustering mean electrode responses (only S and N, importantly) using k-medoids (k = 3) with a correlation-based distance. Shading of the data matrix reflects normalized high-gamma power (70–150 Hz). **c)** Average timecourse by cluster. Shaded areas around the signal reflect a 99% confidence interval over electrodes (n = 99, n = 61, and n = 17 electrodes for Cluster 1, 2, and 3, respectively). Clusters 1-3 showed a strong similarity to the clusters reported in Fig. 2. **d)** Mean condition responses by cluster. Error bars reflect standard error of the mean over electrodes. **e)** Electrode responses visualized on their first two principal components, colored by cluster. **f)** Anatomical distribution of clusters across all participants (n = 6). **g)** Robustness of clusters to electrode omission (random subsets of electrodes were removed in increments of 5). Stars reflect significant similarity with the full dataset (with a p threshold of 0.05; evaluated with a one-sided permutation test, n = 1000 permutations; Methods). Shaded regions reflect standard error of the mean over randomly sampled subsets of electrodes. Relative to when all conditions were used, Cluster 2 was less robust to electrode omission (although still more robust than Cluster 3), suggesting that responses to word lists and Jabberwocky sentences (both not present here) are particularly important for distinguishing Cluster 2 electrodes from Cluster 1 and 3 electrodes.

https://doi.org/10.1038/s41562-024-01944-2

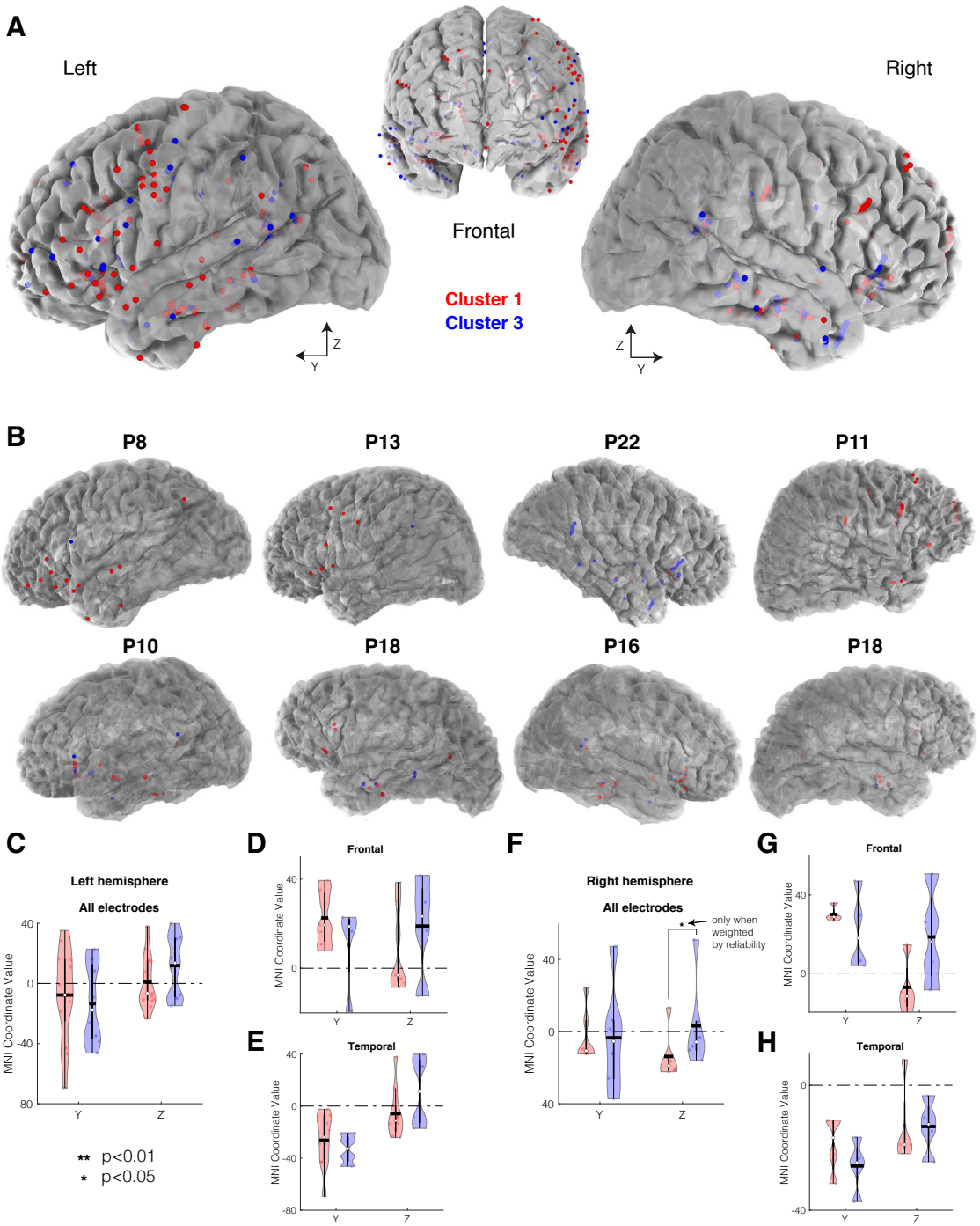


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41562-024-01944-2

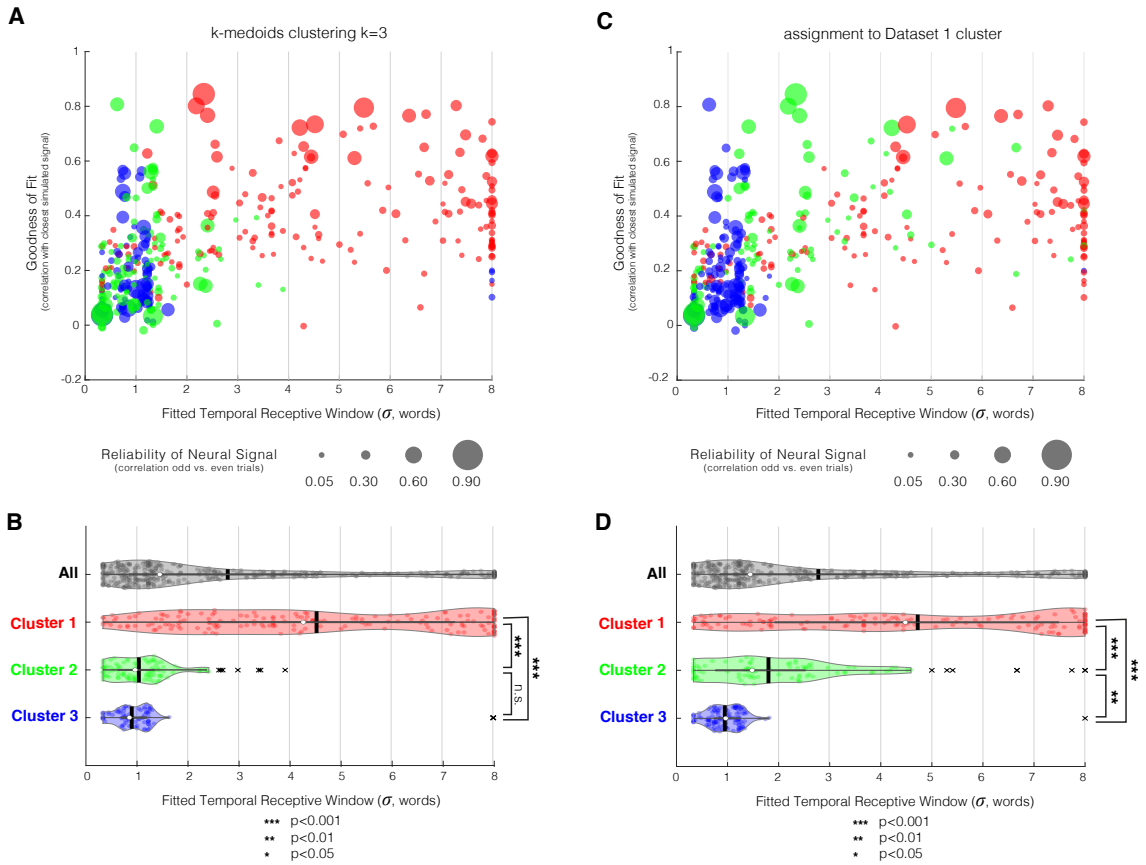
Extended Data Fig. 8 | Dataset 2 electrode assignment to most correlated Dataset 1 cluster under 'winner-take-all' (WTA) approach. a) Assigning electrodes from Dataset 2 to the most correlated cluster from Dataset 1. Assignment was performed using the correlation with the Dataset 1 cluster average, not the cluster medoid. Shading of the data matrix reflects normalized high-gamma power (70–150 Hz). **b)** Average timecourse by group. Shaded areas around the signal reflect a 99% confidence interval over electrodes (n = 142, n = 95, and n = 125 electrodes for groups 1, 2, and 3, respectively). **c)** Mean condition responses by group. Error bars reflect standard error of the mean over electrodes (n = 142, n = 95, and n = 125 electrodes for groups 1, 2, and 3, respectively, as in **b**). **d)** Electrode responses visualized on their first two principal components, colored by group. **e)** Anatomical distribution of groups across all participants (n = 16). **f-g)** Comparison of cluster assignment of electrodes from Dataset 2 using clustering vs. winner-take-all (WTA) approach. **f)** The numbers in the matrix correspond to the number of electrodes assigned to cluster y during clustering (y-axis) versus the number electrodes assigned to group x during the WTA approach (x-axis). For instance, there were 44 electrodes that were assigned to Cluster 1 during clustering but were 'pulled out' to Group 2 (the analog of Cluster 2) during the WTA approach. The total number of electrodes assigned to each cluster during the clustering approach are shown to the right of each row. The total number of electrodes assigned to each group during the WTA approach are shown at the top of each column. N = 362 is the total number of electrodes in Dataset 2. **g)** Similar to **F**, but here the average timecourse across all electrodes assigned to the corresponding cluster/group during both procedures is presented. Shaded areas around the signals reflect a 99% confidence interval over electrodes.

https://doi.org/10.1038/s41562-024-01944-2



Extended Data Fig. 9 | Anatomical distribution of the clusters in Dataset 2. a) Anatomical distribution of language-responsive electrodes in Dataset 2 across all subjects in MNI space, colored by cluster. Only Clusters 1 and 3 (those from Dataset 1 that replicate to Dataset 2) are shown. **b)** Anatomical distribution of language-responsive electrodes in subject-specific space for eight sample participants. **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 (n = 16 participants) represent the mean of all coordinate values for a given participant and cluster. The mean across participants is plotted with a black horizontal line, and the median is shown with a white circle. Vertical thin black boxes within violins plots represent the upper and lower quartiles. Significance is evaluated with a LME model (Methods, Supplementary Tables 3 and 4). The Cluster 3 posterior bias from Dataset 1 was weakly present but not statistically reliable.

https://doi.org/10.1038/s41562-024-01944-2



Extended Data Fig. 10 | Estimation of temporal receptive window (TRW) sizes for electrodes in Dataset 2. As in Fig. 4 but for electrodes in Dataset 2. **a)** Best TRW fit (using the toy model from Fig. 4) for all electrodes, colored by cluster (when k-medoids clustering with k = 3 was applied, Fig. 6) and sized by the reliability of the neural signal as estimated by correlating responses to odd and even trials (Fig. 6c). The 'goodness of fit', or correlation between the simulated and observed neural signal (Sentence condition only), is shown on the y-axis. **b)** Estimated TRW sizes across all electrodes (grey) and per cluster (red, green, and blue). Black vertical lines correspond to the mean window size and the white dots correspond to the median. 'x' marks indicate outliers (more than 1.5 interquartile ranges above the upper quartile or less than 1.5 interquartile ranges below the lower quartile). Significance values were calculated using a linear mixed-effects model (comparing estimate values, two-sided ANOVA for LME, Methods, see Supplementary Table 8 for exact p-values). **c-d)** Same as **A** and **B**, respectively, except that clusters were assigned by highest correlation with Dataset 1 clusters (Extended Data Fig. 8). Under this procedure, Cluster 2 reliably separated from Cluster 3 in terms of its TRW (all ps<0.001, evaluated with a LME model, Methods, see Supplementary Table 9 for exact p-values).

nature portfolio

Corresponding author(s): Tamar Regev, Colton Casto, Evelina Fedorenko

Last updated by author(s): Jun 14, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Intracranial recordings were synchronized with stimulus presentation and stored using the BCI2000 software platform (v3, Schalk et al., 2004)
Data analysis	Matlab 2021a was used for data analysis. Code used to conduct analyses and generate figures from the preprocessed data is available publicly on GitHub at https://github.com/coltoncasto/ecog_clustering_PUBLIC. The VERA software suite used to perform electrode localization can also be found on GitHub at https://github.com/neurotechcenter/VERA.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Preprocessed data, all stimuli and statistical results, and selected additional analyses are available on OSF at https://osf.io/xfbr8/. Raw data may be provided upon request to the corresponding authors and institutional approval of a data-sharing agreement.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	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. Dataset 2: Electrophysiological data were recorded from intracranial electrodes in 16 participants (4 female, aged 21-66 years) with intractable epilepsy.
Population characteristics	Participants (aged 18-66) all had pharmacologically intractable epilepsy and underwent temporary implantation of subdural electrode arrays and depth electrodes to localize epileptogenic foci 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).
Recruitment	Patients with a verbal IQ score >70, as defined by the Wechsler Abbreviated Scale of Intelligence-Second Edition (WASI-II), and with general verbal proficiency in English, as qualitatively evaluated by the experimenters collecting the data, were eligible to participate in the study. All participants gave informed written consent and were not compensated for their participation. The administration of the task was prioritized in patients with left hemisphere frontal and temporal coverage.
Ethics oversight	The Institutional Review Boards at each of the relevant sites (Albany Medical Center (AMC), Barnes-Jewish Hospital (BJH), Mayo Clinic Jacksonville (MCJ), and St. Louis Children's Hospital (SLCH)) approved the protocol—protocol numbers #2061 (AMC), #18-011810 (MCJ), and #201102222 (BJH, SLCH).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was not predetermined by power analysis. Instead, we used all available clinical data from consenting participants during the time period of data collection.
Data exclusions	Dataset 1 (N=6); 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 (N=16); Two further participants were tested but excluded from analyses due to the lack of any language-responsive electrodes (see Language-Responsive Electrode Selection).
Replication	We initially analyzed only Dataset 1. Then, Dataset 2 was analyzed for the purpose of replicating the effects found for Dataset 1, though note that Dataset 2 was not explicitly collected with the intent of replicating our clustering results from Dataset 1. As a result, the experimental design in Dataset 2 had fewer conditions, as the original goal with these data was to simply identify electrodes with a stronger response to sentences versus nonword lists. The principle finding partly replicated: Two out of the three reported response profiles were significantly evident in Dataset 2, and the third profile exhibited a qualitative similarity across the dataset, but this similarity did not reach statistical significance. No additional attempts of replication were made.
Randomization	Stimulus presentation order was randomized across participants.
Blinding	Blinding was not relevant to our study since we did not divide participants into experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging