Peer Review Information

Journal: Nature Human Behaviour
Manuscript Title: Neural populations in the language network differ in the size of their temporal receptive windows
Corresponding author name(s): Tamar I. Regev, Colton Casto and Evelina Fedorenko

Reviewer Comments & Decisions:

Decision Letter, initial version:

3rd May 2023

Dear Dr Regev,

Thank you once again for your manuscript, entitled "Neural populations in the language network differ in the size of their temporal receptive windows", and for your patience during the peer review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of considerable potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

As well as addressing all of the other reviewer comments, we believe that it will be important to carry out the interesting additional analyses suggested by Reviewer #2 (e.g. under "Specifically, how different are responses across trials but within a given condition...") and Reviewer #3 (e.g. "This whole analysis could be significantly enhanced...")

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript, we would hope to receive it within 4 months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

• Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along

with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,
[REDACTED]

REVIEWER COMMENTS:

Reviewer #1:
Remarks to the Author:
This ECoG/sEEG study is a re-analysis of data first presented by Fedorenko et al. (2016) with validation on a second, larger data set using a similar experimental design. Both experiments contained trials with serial visual presentation of words or pseudowords. The trials differed by condition in terms of linguistic structure or semantic content. A yes/no memory probe was presented after each trial. For Dataset 1, the conditions were: Sentences (S), word lists without syntactic structure (W), jabberwocky sentences (J; content words replaced by pseudowords), and lists of pseudowords (N). For Dataset 2, the conditions were S and N. The neural signals analyzed were trial-locked, high-gamma-envelope time series at individual electrodes. Analyses were restricted to 'language-responsive' electrodes defined as a significantly greater response to S than N. Whereas Fedorenko et al. (2016) averaged high gamma activity within individual words/nonwords, the present study examined timecourses across an entire trial of words/nonwords (8 per trial). Unsupervised learning (k-medoids clustering) was used to reveal groups of electrodes with different fine-grained temporal response properties, argued to be related to different temporal receptive windows (TRWs) within the cortical language system. In fact, three such clusters were identified: One with long (~6-8 word) TRWs that activated most to sentences, one with medium (~2-5 word) TRWs that activated as S>W>J>N with notably more activation in S/W, and one with short (~1-2 word) TRWs that activated as S>W>J>N. Clusters also differed in terms of the extent to which time series were locked to individual stimuli (Cluster 1 < Cluster 2 < Cluster 3), and Clusters 2/3 were spread widely over frontotemporal language areas while Cluster 3 was more restricted to posterior language regions. Efforts were made to validate the clustering outcome using an odd-even trial split in dataset 1 and a test of cluster stability using only the S/N conditions across Datasets 1 and 2.

The key conclusions were: (i) there are three distinct temporal profiles across-language responsive electrodes; (ii) the key feature distinguishing these profiles is the length of their TRWs (corresponding roughly to sentence-, phrasal-, and word-level analysis windows), and (iii) these response profiles are highly distributed across the frontotemporal language network.

This a very well-written and thorough manuscript that makes a significant contribution relative to the existing literature by generating a plausible argument for three discrete TRWs within the language system that appear to be spatially distributed across language regions, suggestive of hierarchical processing systems that operate in parallel. I found the Future Directions section to be quite well developed and I agree that this line of work has potential to generate important follow up studies likely to advance our neurocomputational understanding of cortical language systems. However, I have some concerns related to interpretation of the data, as well as several methodological concerns, that I would like the authors to

respond to (continue below signature block).

Respectfully,
Jonathan H. Venezia
VA Loma Linda Healthcare System

1. Heterogeneity/replicability of responses within Cluster 2. While the differences between Clusters 1 and 3 appear robust, the functional significance of Cluster 2 relative to 1/3 is difficult to ascertain. While high reliability electrodes in Cluster 2 do seem to separate well from those in Clusters 1/3 in a low-dimensional space (Fig. 3C), the spread among high reliability electrodes within Cluster 2 is – to visual approximation – larger. This is particularly evident from the TRW modeling, where Cluster 2 electrodes clearly have the broadest distribution (Fig. 4C), even if we restrict to high reliability electrodes (Fig. 4B, large circles). In the manuscript, it is argued that Cluster 2 reflects processing of short phrases, evidenced by a buildup of activity over 2-3 words, followed by a plateau in the S condition and a decline in the other conditions (Fig. 3D/E), the latter reflecting that Cluster 2 electrodes essentially "give up" on phrase building when it becomes clear that no such structure is present in the stimulus. It is further argued that electrodes in Cluster 2 do not respond to linguistic structure independent of word meaning because the response profiles in S and J are different. However, much like the S response, the response to J in the Cluster 2 medoid (Fig. 3D) appears to increase for 2-3 words and then plateau (though at a lower overall amplitude). Moreover, the response to J in Cluster 2 exceeds the response to N, which seems consistent with a response to abstract structure (acknowledging the potential confound that J includes at least some real function words). The fact that Cluster 2, in particular, fails to replicate strongly in Dataset 2 is suspicious given that the S and N conditions seem to contrast most strongly in terms of their potential to produce phrasal structure at any level. Indeed, the authors acknowledge that the inclusion of the W and J conditions may be important to revealing Cluster 2. The same can be seen in Fig. S3E, where Clusters 1/2 overlap much more strongly in the low-D space when clustering is restricted to S and N only in Dataset 1 (see also Fig. S7, where 'forcing' a fit with Cluster 2 in Dataset 2 yields TRWs that now overlap strongly with Cluster 1). My point here is that there appears to be a good deal of functional heterogeneity within Cluster 2, potentially related to experimental condition (see a related explanation below), which challenges the explanation that Cluster 2 simply reflects combinatorial processing across words at roughly a phrasal scale. Indeed, the reliable presence of Cluster 2 is critical to broader claims about functional heterogeneity and distributed processing with respect to TRWs.

2. Clustering approach.

(a) k-medoids vs. k-means. It is argued that k-medoids is preferred over k-means because the use of a correlation distance metric rules out cluster differences based on scale (as opposed to shape) of the clustered timecourses. However, since the mean timecourses for S/W/N/J were concatenated for each electrode prior to clustering, and since there tended to be differences in activation amplitude by condition (Figure 3F), correlations between electrodes could have been influenced by scale due to Simpson's paradox (i.e., correlations were obtained across different "groups" of time points – the conditions – whose means were potentially different). One possible solution here would be to z-score the timecourses within condition prior to concatenation. Follow-up analyses based on TRW modeling and degree of stimulus locking assuage this concern somewhat, though I remain concerned that condition differences in overall amplitude may be an important factor allowing Cluster 2 to be distinguished from Clusters 1 and 3, given my above concern about heterogeneity in TRWs within Cluster 2.

(b) Electrode reliability. Arguably, electrode reliability (odd-even trials correlation) should be included in the clustering algorithm from the outset by using a weighted k-medoid (e.g., R package WeightedCluster) or k-means approach. This could improve the separability of the clusters given that, as shown in Fig. 3C, low-reliability electrodes tend to separate poorly. Further, given that k clustering is a "hard" clustering approach, distances from the cluster medoids should potentially be used as weights in subsequent LME analyses for which 'cluster' is used as a categorical variable. For example, this may have a large impact on whether the data support spatial separation of the clusters because the best exemplars (e.g., of Clusters 1 and 2) may be more well separated than the poor exemplars. With respect to spatial separation specifically, an alternative would be to employ clustering methods capable of considering the spatial proximity of the electrodes in the clustering solution (e.g., R package hclustgeo). Spatial constraint could be considered bias

the clustering solution toward a spatially segregated outcome, but it would be useful to know if such an outcome matches, e.g., existing patterns from fMRI, which are being challenged here.

3. Binary classifier. The authors place much weight on the fact that activation in Cluster 1 is able to discriminate S from W in the earliest time bins, while this takes 2-3 words to build up in Cluster 2. This is taken to reflect the time it takes Cluster 2 electrodes to "rule out" the existence of phrasal structure in W. However, unless I'm interpreting Fig. 5A incorrectly, it seems that Cluster 2 is reliably able to discriminate S from W in less than 500 ms, roughly the time it takes to present a single word in the fast-timing paradigm (450 ms). More generally, I assume "reliable discrimination" is taken to mean the first time point at which 10-fold cross-validated classifier accuracy differs significantly from chance (50%, one-sample t-test, reflected as stars in Fig. 5A). This analysis is likely too liberal because (i) assumptions of the t-test are often violated when applied to classifier accuracy (e.g., non-Gaussianity); (ii) the analysis does not account for temporal autocorrelation in the underlying time courses and interdependence between time bins introduced by overlapping windows (100 ms, 50% overlap); and (iii) there was no correction for multiple comparisons across time bins. In this sense, the classifier is very much a 1-D analog of MVPA in fMRI. Stelzer et al. (2013, Neuroimage) address the problems related to group-level inference on cross-validated classification accuracies. Short of their recommended approach, I would recommend at least adopting non-parametric approach estimation of the null distribution of group-level accuracy combined with multiple comparison correction across bins. The broader point though, is that inferences drawn with respect to when activation between conditions begins to differ are highly dependent on the statistical criterion. Moreover, little attention is paid to other aspects of the data, e.g., that classifier accuracy is much higher overall in the W vs. N contrast for Cluster 2 than Cluster 1. Finally, the J vs. N contrast (currently omitted) could be highly relevant here, given above comments related to J re: Cluster 2.

4. LME models. All three LME models (stimulus locking, TRW, and MNI coordinates across electrodes and participants) include a random effect of cluster within participant, where cluster membership is dummy coded as a categorical variable. It is unclear to me how the random effect of cluster was estimated given that all levels of cluster (1, 2, 3) were not present within all levels of participant. That is, Cluster 3 electrodes were present in only 4/6 participants (p. 7, but visually appears to be 5/6 in Figure 6B), which would result in a vector of all zeros for some participants in the dummy coded variable column. Were the participants without Cluster 3 excluded from the LMEs? If not, is there evidence that the estimates of participant-level variance were robust and valid? The denominator DFs in the reported LME tables also seem high. If no corrective approximation (e.g., Satterthwaite, available in MATLAB fitlme) was applied during the computation of the ANOVA, justification is required as this could severely influence the threshold for significance.

Taken together, these issues primarily raise concerns related to (i) whether the three clusters identified correspond to three unique mechanisms primarily in terms of TRWs in the language system, or whether some other aspect of processing related to experimental condition plays a crucial role; and (ii) whether claims about distributed processing, derived from the spatial distribution of cluster-labeled electrodes, are supported robustly by the data.

Minutiae:

I would like to see the behavioral results reported for both datasets. I found the results for four participants from Dataset 1 in in Fedorenko et al. (2016), but full reporting for Dataset 1 and Dataset 2 would be welcome.

p. 24: "The following electrodes (electrodes; we use these terms interchangeably)"

p. 25, first line: A couple details should be added on how time warping was performed

p. 28, first line of "Electrode discrimination between conditions" section: examined ◊ examine

p. 28: "amplitude and offset of the sinusoidal by searching parameter combinations." Sinusoidal should be either "sinusoidal function" or "sinusoid"

p. 30: "see 1 above," I can figure out what this is referring to but not sure what '1' means

Reviewer #2:
Remarks to the Author:
SUMMARY

The authors recorded intercranial EEG data from patients who completed a visual reading task. Four experimental conditions were employed: (i) full sentences; (ii) scrambled word lists; (iii) jabberwocky; (iv) non word lists. Cluster analyses reveal that there are three main sources of temporal variance in the data: one cluster which increases magnitude as a function of sentence duration, and is selective for sentences; another which responds equally to sentences and word-lists; and a third which responds to each individual written stimulus regardless of stimulus condition or location in the sequence. The authors interpret this result as indicating that different neural populations are governed by 'temporal receptive windows' of different sizes. Localization of these electrodes suggest that neural populations with different window sizes are intermixed, and not organized by gross anatomical region.

Overall the paper is clear and well written. I have some concerns that I outline in detail below. The most major of which relate to the specificity of the claims and some of the analytical approaches

MAJOR

[1] Vague interpretation

One difficulty I am having is understanding the broader implication of this result, and how it informs neural language processing. I think this is partly because the differences across stimuli are fully ignored, which makes it very difficult to link the difference in physiological responses across electrodes to the actual computation that is implemented. The authors do mention this shortcoming of analyzing responses "regardless of their structure or meaning", but their suggestion for overcoming this is to use single neuron recordings and deep learning language models. The data being analyzed here are some of the most powerful and valuable that the human neuroscience community currently has access to, and I believe that it is possible to address this shortcoming with the current dataset.

Specifically, how different are responses across trials but within a given condition? What explains that variance? Are all words processed the same, or is there variability? Does the "build-up" in activity have the same gradient across sentences? Is the gradual build-up actually there on all sentences or is it more jagged on an individual trial basis? To me, these are all important questions that would need to be answered in order to fully interpret the results. The clustering approach is useful for identifying different types of electrode dynamics but that seems like just the first step to actually understanding what those dynamics mean. The temporal resolution of the intercranial data are not being taken full advantage of, due to the averaging across different stimuli.

[2] Modeling approach

In a similar vein, I had some concerns regarding the modeling. I could not find a description of what hypothesized neural generators were being modeled here, and what the underlying hypothesis is of what processes are actually happening. This felt more like a way of capturing the dynamics of the already observed time courses rather than a prior conceived hypothesis. That is fine, but it needs to be discussed as such. And, framed as an analysis method not a model, if that is the goal.

Building on that, the authors use a gaussian kernel with variable widths to estimate the different window sizes. These kernels are centered around the onset of each word, by virtue of the convolutional method. This means that the neural activity is assumed to increase prior to the word being presented on the screen, and the increase is assumed to be symmetric with the decrease. Why are these assumptions built into the model? In a few places, the authors refer to these processes as oscillatory. If that is their mechanistic assumption, then the process should be modeled as such, with sinusoids rather than cumulative gaussian

kernels. Otherwise, it should be modeled as a series of evoked responses with a neurophysiologically realistic response function, and a hypothesis for the underlying mechanistic generator. In the most ideal case, this would be performed on, and be able to account for variance in, the single trial data.

[3] Cluster-electrode assignment

The paper relies heavily on assigning electrodes to clusters. It would be useful to understand how "pure" an electrode is relative to its cluster assignment. Does it also have significant weight on the other clusters? How much of the other timecourse dynamics are intermixed within an electrode response? Given that the results suggest that the spatial organization is essentially random, it would seem likely that a single electrode would capture responses from neural populations that also contain different dynamics. Understanding to what extent this is the case would be helpful for better interpreting the heterogeneity of the results.

MINOR

* I found the motivation for using invasive recordings a little surprising, in terms of claiming that fMRI has poor spatial resolution. Certainly the temporal resolution is better for neurophysiological recordings, but the blanket statement that fMRI has poor spatial resolution was puzzling. What about high Tesla systems for example? In addition, the electrodes being used here are a couple of millimetres in diameter with a spacing of 1cm — this is not such excellent spatial resolution within a single subject. I would recommend that the authors remove entirely the claim about spatial resolution differences.

*The authors seem to be heavily citing their own work, and missing out important and relevant research from other research labs.

* The authors do a good job of demonstrating that the clusters they find are robust across exclusive partitions of the data. And I understand that the authors motivate the use of 3 clusters by use of the "elbow" method. I also appreciate adding different 'k' values in the supplementary. However, the amount of variance accounted for by each of the first 10 clusters may still be interpretable, beyond the fourth cluster. It would be useful to include visualization of each of the 10 clusters similar to that in Figure 3D.

* The use of a [1, -1] ideal label assignment for the sentence vs. nonword lists, in order to identify language responsive electrodes, is maybe not the best, given that the high gamma signal can never be below zero. It might be good to check that an assignment of [1, 0] gives the same result.

* Minor thing but I recommend avoiding saying "an electrode processes information" and instead more precisely talk about neural populations as captured by the electrode

Reviewer #3:
Remarks to the Author:
The study describes an analysis of the temporal receptive windows of different neural populations which respond to language by comparing the responses of these populations to stimuli with different amounts of linguistic content. Overall, the study's main finding is interesting and enhances our understanding of the distribution of neural populations with different temporal integration periods. The authors also show good robustness for this finding across datasets. However, the current analyses focus too much on clusters rather than on individual electrode properties inside each cluster. I have few major concerns:

1. The TRW modeling analysis is intuitive, but it doesn't add much in its current state because the average time courses of each cluster seem to show the same information already because the analysis only discusses the results at the cluster level, rather than the individual electrode level. Cluster 1 and cluster 2 have very wide distributions of fitted TRW (fig 4C), and it would have been nice to analyze this more and look at the electrodes in these groups that were on the far ends of the distribution to look at what about them put them in their respective cluster despite the TRW model choosing a very different receptive field. This whole analysis could be significantly enhanced if, rather than just looking at each cluster, each

electrode's best fitted receptive window was used to color a visualization of the electrodes on the brain, which may help show some interesting gradients of TRW along different processing pathways. While the authors seem to claim that these clusters of population responses illustrate discrete types of responses throughout the language network, that doesn't mean that the TRW of populations doesn't follow a more continuous gradient.
2. The following argument needs to become more compelling; since the more-reliable (i.e. less noisy) electrodes show clearer separation among clusters in low-dimensional space, then this is evidence that these clusters characterize the structure of responses and argue against a uniform continuum of response types. If the 2D Principal components plot (fig 2C) shows the less reliable electrodes towards the center, couldn't that be because the less reliable electrodes have an average time course which gets washed out more after averaging, so the principal components go towards the average more, rather than towards the clusters? If so, this figure doesn't seem like it's evidence that there isn't a continuum, because this should naturally happen to less reliable electrodes. If these principal components come from average time-courses, then they are naturally tied to the reliability metric of correlation between even and odd trials.
3. Another critical factor is the issue of subject variability. While the same clusters can appear in different subjects, this is not enough to claim that all subjects have the same response property. For example, one can see from Figure 3B that cluster 1 only shows a clear incremental response to sentences in the top subject. While this figure indicates that cluster 1 also appears in different subjects, it is not clear from this figure if other subjects (red rows) show the same incremental pattern (at least not from 3B). Again, the problem with clustering is that this analysis forces a categorical label upon electrodes, based on a comparison of a similarity metric. However, if an electrode belongs to a cluster, it doesn't necessarily mean that electrode does the same thing as the mean of the cluster. What would be helpful is to include the average cluster plots (3E) for individual subjects and show they all have the same response pattern across the four stimuli.

Minor points:
1. Please add more details about the stimuli. I understand that stimulus information is available in previous studies, but this paper needs to be self-contained.
2. The citations are pretty narrow and rely heavily on self-referencing. This needs to be expanded. Several relevant recent papers have looked at similar questions and showed a similar distributed pattern of linguistic processing, including Ding et al. 2016, de Heer et al. 2017, Keshishian et al. 2023, and Brodbeck et al., 2018.
3. I do not think the first figure is important enough to be included as main and can be moved to supplemental, discussion, or removed entirely. The paper only uses ECoG, which has been used for decades to study speech in humans. There is no need to have a main figure that discusses its superior spatial resolution compared to other recording techniques.

Author Rebuttal to Initial comments

REVIEWER COMMENTS:

Reviewer #1:

[summary omitted]

This a very well-written and thorough manuscript that makes a significant contribution relative to the existing literature by generating a plausible argument for three discrete TRWs within the language system that appear to be spatially distributed across language regions, suggestive of hierarchical processing systems that operate in parallel. I found the Future Directions section to be quite well developed and I agree that this line of work has potential to generate important follow up studies likely to advance our neurocomputational understanding of cortical language systems.

Thank you for the positive evaluation of this work and for all the helpful comments below.

However, I have some concerns related to interpretation of the data, as well as several methodological concerns, that I would like the authors to respond to.

1. **Heterogeneity/replicability of responses within Cluster 2**. While the differences between Clusters 1 and 3 appear robust, the functional significance of Cluster 2 relative to 1/3 is difficult to ascertain. While high reliability electrodes in Cluster 2 do seem to separate well from those in Clusters 1/3 in a low-dimensional space (Fig. 3C), the spread among high reliability electrodes within Cluster 2 is – to visual approximation – larger. This is particularly evident from the TRW modeling, where Cluster 2 electrodes clearly have the broadest distribution (Fig. 4C), even if we restrict to high reliability electrodes (Fig. 4B, large circles).

You are right that electrodes in Cluster 2 show a higher spread / greater variability, both in the PCA space (**Figure 2C** in the revised manuscript) and in the estimated TRW values (**Figure 4B, C**). It is worth noting though that the spread in the PCA space is not much larger for Cluster 2 than for Cluster 1 electrodes; and for the estimated TRW values, the variability is greatly reduced if we restrict ourselves to higher-reliability electrodes (larger circles). That said, we have created a visualization of single electrode time-courses in order to both increase transparency of single-electrode data, as well as gain more insight into how this variability across Cluster 2 electrodes manifests at the level of the neural signals themselves (**Figure R1** below; included as **Figure S4** in the revised manuscript).

In particular, **Figure R1** displays the neural signals in all electrodes (averaged over trials) within each cluster (**A**: Cluster 1; **B**: Cluster 2; **C**: Cluster 3) *sorted by their distance to the cluster medoids*, with electrodes toward the bottom of each image being closest to the medoids. The electrodes are colored by the reliability of the signals (estimated by correlating responses to odd and even trials), with more reliable timecourses in pink, and we display for each electrode the estimated TRW to the left of each timecourse. This visualization makes apparent that **many individual response profiles look visually similar to the prototypical cluster response profile** for all clusters, including Cluster 2. As the distance from the medoid increases, the electrode responses are less reliable, and the dynamics of the signal become less interpretable.

This visualization also highlighted that a couple of highly-reliable electrodes exhibit response profiles that are unlike the prototypical cluster response profiles. For example, an electrode in Cluster 2 responds strongly *only* to the first word/nonword in all conditions, and an electrode in Cluster 3 shows word-locked responses for all words, *except the first word*. These patterns do not fall within the TRW framework. These electrodes highlight that the clusters that we uncovered may not be the *only* responses that exist within the language network, as we have now acknowledged (Results). However, the rarity of these responses also emphasizes that the clusters that we uncovered do seem to capture a substantial amount of functional heterogeneity of response patterns in the language network. Thank you for encouraging us to explore within-cluster timecourse variability.

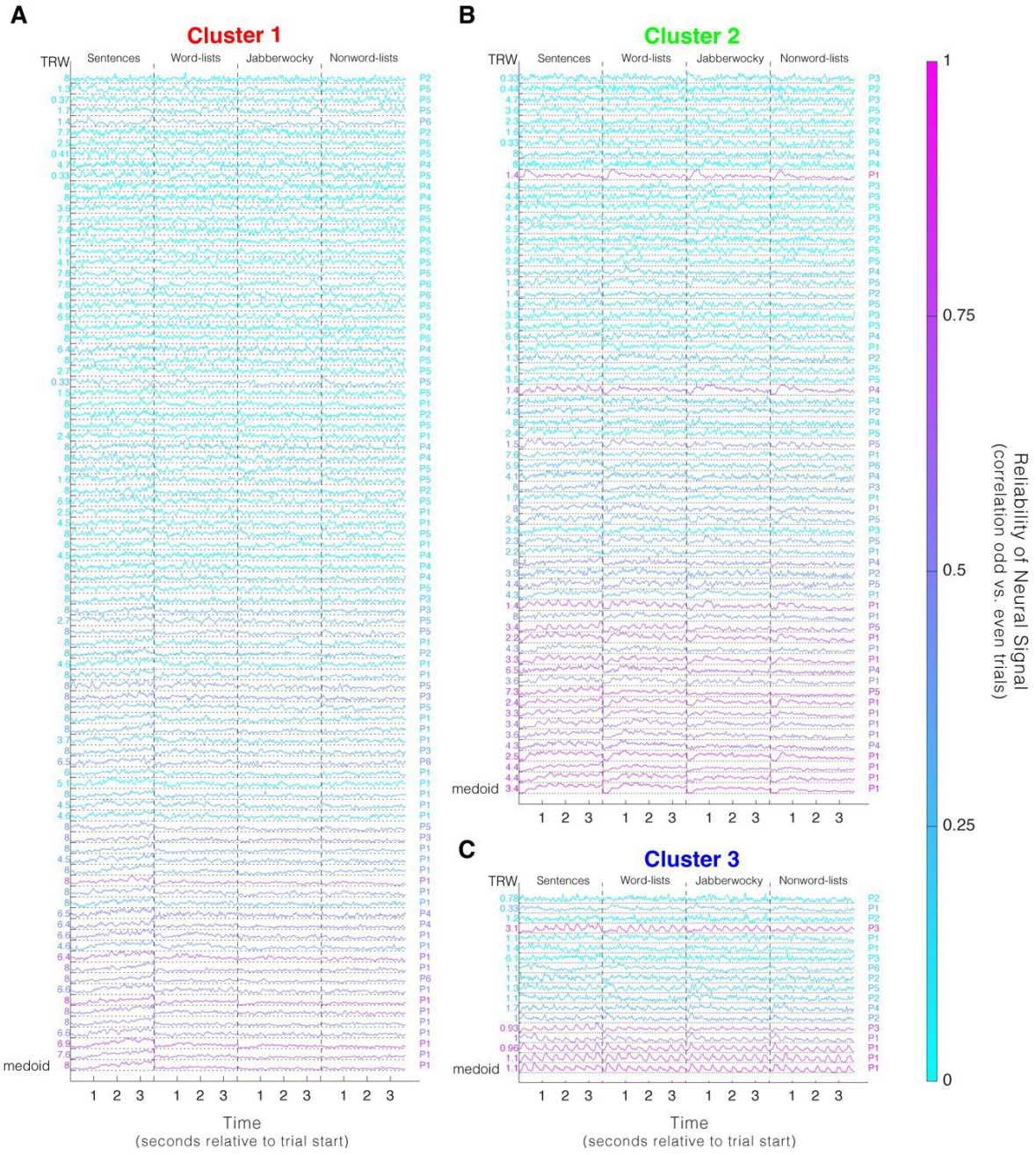


Figure R1 (Figure S4 in Supplementary Information). **A-C**) All Dataset 1 electrode responses. The timecourses (concatenated across the four conditions, ordered: Sentences, Word-lists, Jabberwocky, Nonword-lists) of all electrodes in Dataset 1 sorted by their correlation to the cluster medoid (shown at the bottom of each cluster). Colors reflect the reliability of the measured neural signal, computed by correlating responses to odd and even trials (**Figure 2D**). The estimated temporal receptive window (TRW) using the toy model from **Figure 4** is displayed to the left, and the participant who contributed the electrode is displayed to the right. There is strong consistency in the responses from individual electrodes within a cluster (with more variability in the less reliable electrodes), and electrodes with responses that are more similar to the cluster medoid tend to be more reliable (more pink).

In the manuscript, it is argued that Cluster 2 reflects processing of short phrases, evidenced by a buildup of activity over 2-3 words, followed by a plateau in the S condition and a decline in the other conditions (Fig. 3D/E), the latter reflecting that Cluster 2 electrodes essentially "give up" on phrase building when it becomes clear that no such structure is present in the stimulus. It is further argued that electrodes in **Cluster 2 do not respond to linguistic structure independent of word meaning** because the response profiles in S and J are different.
However, much like the S response, the response to J in the Cluster 2 medoid (Fig. 3D) appears to increase for 2-3 words and then plateau (though at a lower overall amplitude). Moreover, the response to J in Cluster 2 exceeds the response to N, which seems consistent with a response to abstract structure (acknowledging the potential confound that J includes at least some real function words).

Thank you for the opportunity to clarify this point, as we did not phrase our position accurately. We did not mean to say that "*electrodes in Cluster 2 do not respond to linguistic structure independent of word meaning*" in the sense that they are not sensitive to abstract structure. We agree with the reviewer that these electrodes are indeed sensitive to abstract structure as evidenced by their stronger responses to the Jabberwocky condition than the Nonword-lists condition. What we had intended to say is that **no electrode in this cluster** (and, in fact, no electrode in any other cluster) **is sensitive *ONLY* to structure**, as would be evidenced by a profile where the responses are similar in both magnitude and shape between the Sentences and Jabberwocky conditions, with both being higher than the two unstructured conditions (word-lists and nonword-lists). Instead, all electrodes appear to be sensitive to both word meanings *AND* structure, as evidenced by stronger responses to word-lists (which contain word meanings) and Jabberwocky (which contain abstract structure) than nonword-lists (which contain neither word meanings nor abstract structure).

The fact that Cluster 2, in particular, fails to replicate strongly in Dataset 2 is suspicious given that the S and N conditions seem to contrast most strongly in terms of their potential to produce phrasal structure at any level. Indeed, the authors acknowledge that the inclusion of the W and J conditions may be important to revealing Cluster 2. The same can be seen in Fig. S3E, where Clusters 1/2 overlap much more strongly in the low-D space when clustering is restricted to S and N only in Dataset 1 (see also Fig. S7, where 'forcing' a fit with Cluster 2 in Dataset 2 yields TRWs that now overlap strongly with Cluster 1). My point here is that there appears to be a good deal of **functional heterogeneity within Cluster 2**, potentially related to experimental condition (see a related explanation below), which challenges the explanation that Cluster 2 simply reflects combinatorial processing across words at roughly a phrasal scale. Indeed, the reliable presence of Cluster 2 is critical to broader claims about functional heterogeneity and distributed processing with respect to TRWs.

Thank you for this comment. You are right that the W and J conditions do seem helpful for distinguishing Clusters 1 and 2 in Dataset 1. However, examination of the profiles of individual electrodes in Dataset 2 reveals several electrodes with responses to S and N that look very similar to the prototypical responses of Cluster 2 electrodes in Dataset 1. To evaluate the presence of the Cluster 2 profile in Dataset 2, we performed a new analysis that is slightly less conservative than performing unsupervised clustering

"from scratch" but is still able to test for the existence of Cluster-2-like responses. Specifically, instead of performing k-medoids clustering, we assigned each electrode in Dataset 2 to one of 3 groups according to its correlation with the average cluster responses found in Dataset 1 (**Figure R2** below and included as **Figure S8** in the manuscript). This approach revealed response profiles that were ***strikingly similar to the response profiles found in Dataset 1***, including a ***replication of a Cluster-2-like pattern***. This result indicates that Cluster-2-like responses are robustly present in Dataset 2 as well, despite the fact that this profile did not robustly emerge in the most conservative version of the analysis, when Dataset 2 clustering was done "from scratch". Furthermore, under this approach we find a clearer separation of fitted TRW by cluster (panels **A** and **B** vs. **C** and **D** in **Figure R3** below and **Figure S9** of the manuscript).

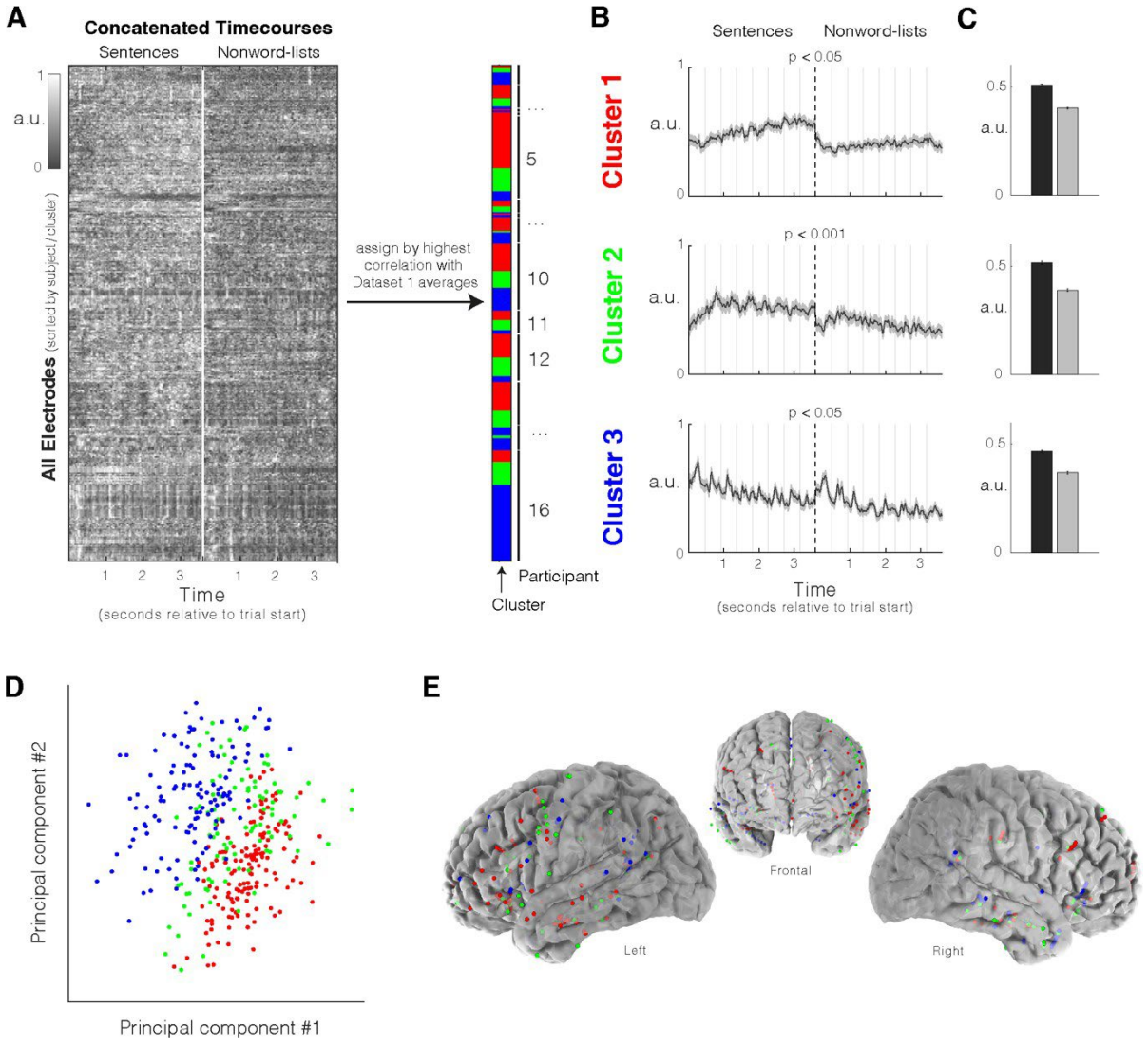


Figure R2 (Figure S8 in Supplementary Information). Dataset 2 electrode assignment to the most correlated Dataset 1 cluster. **A)** Assigning electrodes from Dataset 2 to the most correlated cluster from Dataset 1. Assignment is performed using the correlation with the Dataset 1 cluster average. Shading of the data matrix reflects normalized high-gamma power (70-150Hz). **B)** Average timecourse by cluster. Shaded areas around the

signal reflect a 99% confidence interval over electrodes. Under this procedure, all clusters – including Cluster 2 – are more similar to the clusters in Dataset 1 then would be expected by chance (C1: p=0.025, C2: p=0.001, C3: p=0.029; evaluated with a permutation test, Methods). **C)** Mean condition responses by cluster. Error bars reflect standard error of the mean over electrodes. **D)** Electrode responses visualized on their first two principal components, colored by cluster. **E)** Anatomical distribution of clusters across all participants (n=16).

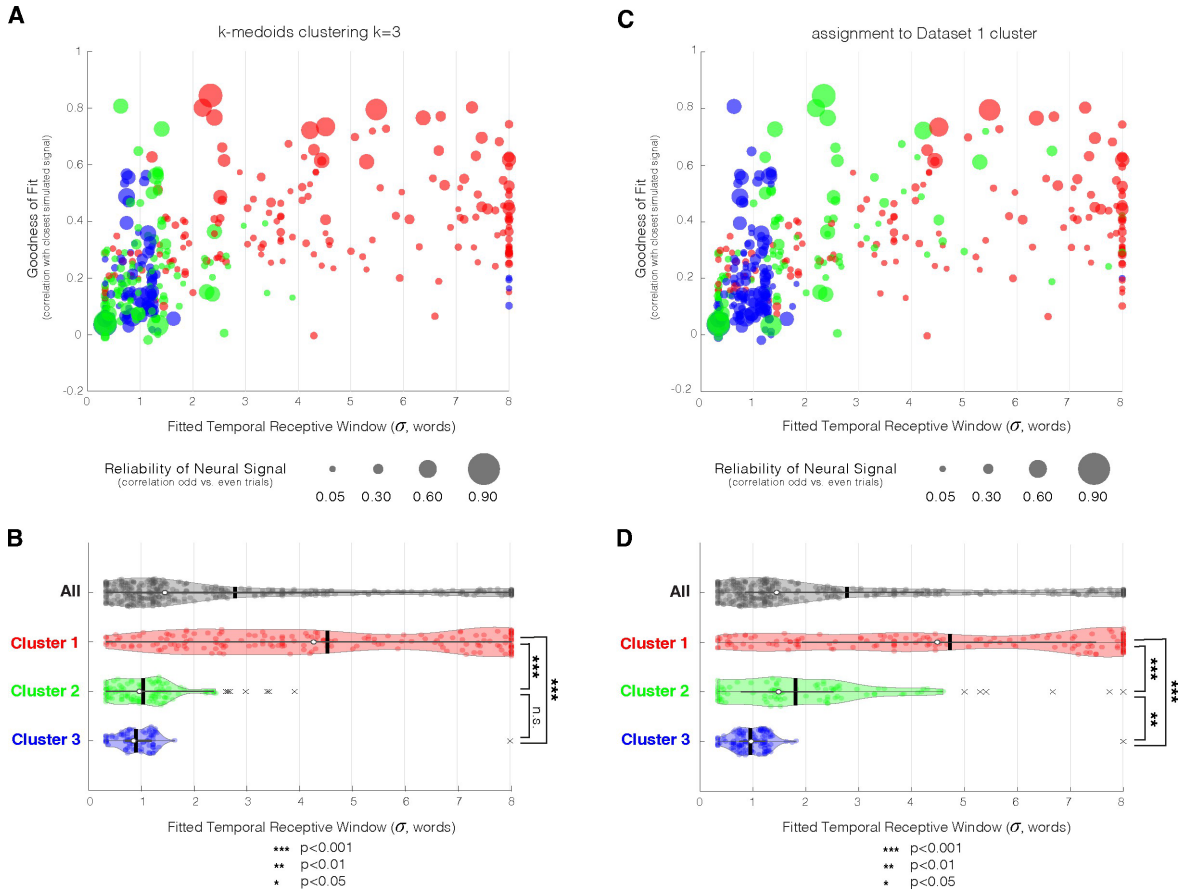


Figure R3 (Figure S10 in Supplementary Information). Estimation of temporal receptive window (TRW) sizes for electrodes in Dataset 2. Same as main-text **Figure 4** but for the electrodes in Dataset 2. **A)** Best TRW fit (using the toy model from **Figure 4**) for all electrodes colored by cluster (when k-medoids clustering with k=3 is applied, **Figure 6**) and sized by the reliability of the neural signal as estimated by correlating responses to odd and even trials (**Figure 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 are calculated using a linear mixed-effects (LME) model (Methods, **Table S8**). **C-D)** Same as **A** and **B**, respectively, except clusters are assigned by highest correlation with Dataset 1 clusters (**Figure S8**). Under this procedure, Cluster 2 reliably separates from Cluster 3 in terms of its TRW (all ps<0.001, evaluated with a LME model, Methods, **Table S9**).

2. Clustering approach.

(a) k-medoids vs. k-means. It is argued that k-medoids is preferred over k-means because the use of a correlation distance metric rules out cluster differences based on scale (as opposed to shape) of the clustered timecourses. However, since the mean timecourses for S/W/N/J were **concatenated** for each electrode prior to clustering, and since there tended to be differences in **activation amplitude by condition** (Figure 3F), correlations between electrodes could have been influenced by scale due to Simpson's paradox (i.e., correlations were obtained across different "groups" of time points – the conditions – whose means were potentially different). One possible solution here would be to z-score the timecourses within condition prior to concatenation. Follow-up analyses based on TRW modeling and degree of stimulus locking assuage this concern somewhat, though **I remain concerned that condition differences in overall amplitude may be an important factor allowing Cluster 2 to be distinguished from Clusters 1 and 3, given my above concern about heterogeneity in TRWs within Cluster 2.**

To clarify: the use of a correlation distance metric ensures that **overall** shifts in amplitude, **across conditions** (e.g., overall voltage scale differences between electrodes), would not affect the results. However, the **relative** amplitude differences among the conditions within an electrode constitute an important feature of the functional response patterns of neural populations that we would not want to eliminate. Our condition-timecourse concatenation approach ensures that *the relative amplitude differences between conditions are fixed per electrode,* and eliminating these differences could distort electrode responses in non-functionally-meaningful ways.

This said, we performed the analysis that you suggested: clustering the signals after z-scoring each condition separately. The results are shown in **Figure R4** below. The general pattern of the clusters remained the same, evidencing robustness to this z-scoring procedure, which demonstrates that the clusters are not characterized solely by the relative magnitude differences among the conditions but also by the *temporal response patterns* within each condition. Because the response patterns are more difficult to interpret without condition differences in amplitude, as well as due to space limitations, we decided to include this analysis on OSF and not in the Supplementary Information.

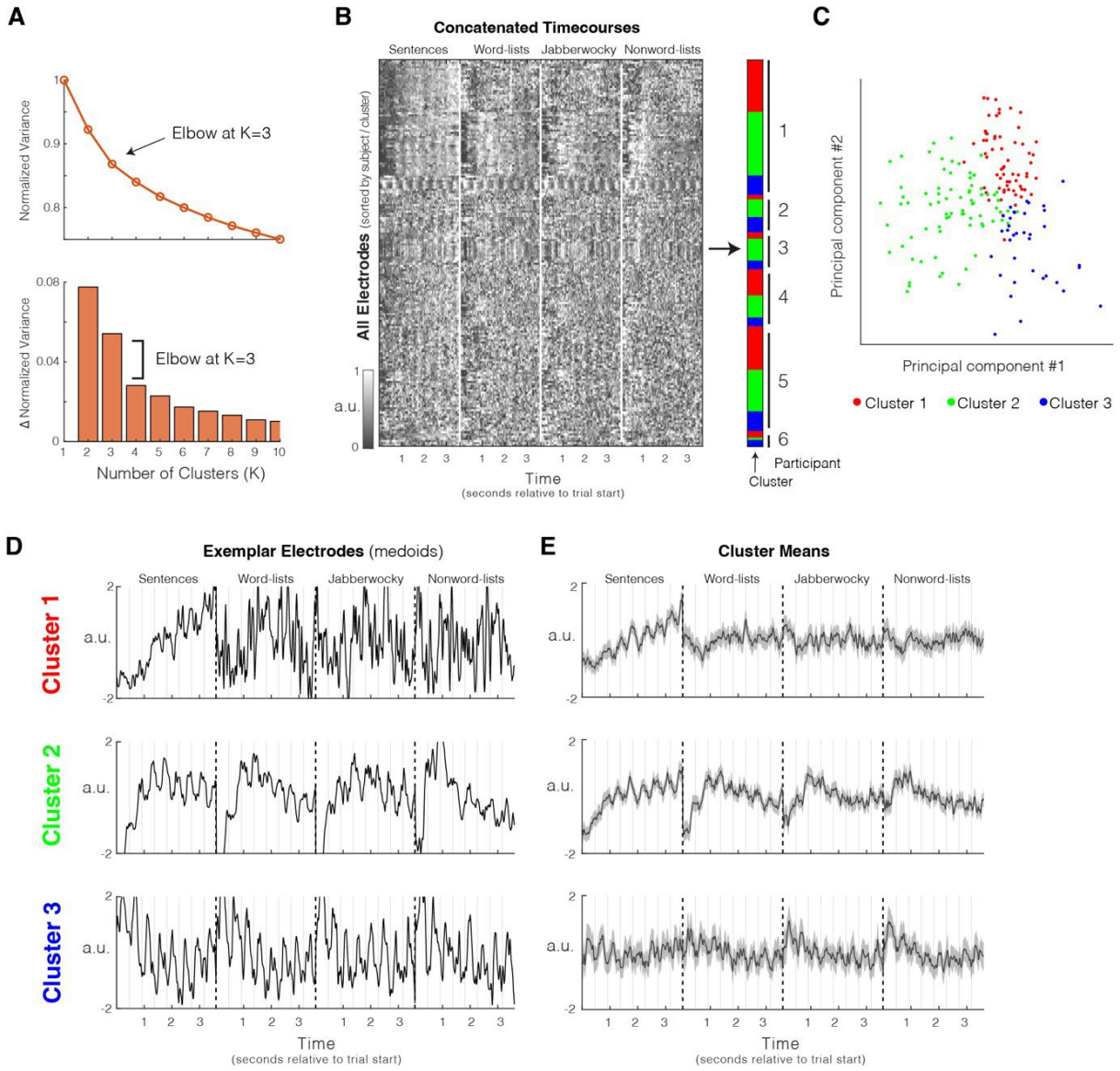


Figure R4. Clustering electrode timecourses after z-scoring the responses per condition. This figure mirrors **Figure 2** in the paper, except that we z-scored the responses for each condition separately, which removes the relative amplitude differences between the four conditions. The same three clusters emerged, demonstrating that the clusters were not characterized solely by the relative magnitude differences among the conditions but also by the *temporal response patterns* within each condition.

(b) Electrode reliability. Arguably, electrode reliability (odd-even trials correlation) should be included in the clustering algorithm from the outset by using a **weighted k-medoid** (e.g., R package WeightedCluster) or k-means approach. This could improve the separability of the clusters given that, as shown in Fig. 3C, low-reliability electrodes tend to separate poorly. Further, given that k clustering is a "hard" clustering approach, **distances from the cluster medoids should potentially be used as weights in subsequent LME analyses for which 'cluster' is used as a categorical variable.** For example, this may have a large impact on whether the data support spatial separation of the clusters because the best exemplars (e.g., of

Clusters 1 and 2) may be more well separated than the poor exemplars. With respect to spatial separation specifically, an alternative would be to employ clustering methods capable of considering the spatial proximity of the electrodes in the clustering solution (e.g., R package hclustgeo). Spatial constraint could be considered bias the clustering solution toward a spatially segregated outcome, but it would be useful to know if such an outcome matches, e.g., existing patterns from fMRI, which are being challenged here.

Thank you for your suggestion to use electrode reliability as a weighting factor. We have performed two new analyses where we take electrode reliability into account.

First, we examined how the clustering is affected by the exclusion of lower-reliability electrodes. To this end, we repeated the clustering analysis several times, while omitting electrodes below a parametrically sampled reliability threshold, down to n=53 (~30%) most reliable electrodes (with reliability >0.4). The results are shown in **Figure R5** below. This analysis strengthened our confidence in k=3 clusters being an optimal number of clusters for this dataset: the elbow became sharper in the subsets of increasingly more reliable electrodes (see also our response to Reviewer #3's second comment). However, for our main analyses, we prefer to use a standard k-medoids approach on all electrodes, as we believe this is the most conservative approach, and the fact that these three response profiles are recoverable without any filtering of electrodes by reliability speaks to their robustness. Because this figure is useful in demonstrating the separability of the clusters, we have now added it as an inset in **Figure 2A** in the revised manuscript.

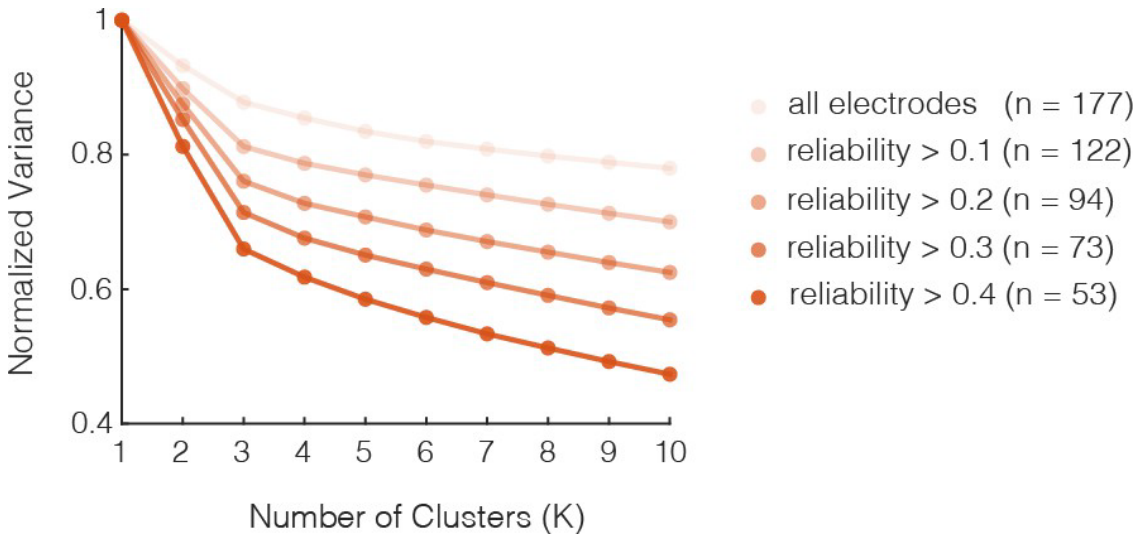


Figure R5 (inset for Figure 2A in the revised manuscript). Clustering while omitting electrodes below a parametrically sampled reliability threshold. Normalized variance is presented against the number of clusters (k), to illustrate the elbow method. The 'elbow' (the point of transition between a steeper and a more moderate slope) gets more pronounced when eliminating lower-reliability electrodes, which strengthens our claim that k=3 best describes these data.

Second, as you suggested, we re-ran the LME models that test for spatial effects while adding a reliability-based weighting factor. The addition of the weighting factor did not affect the results that we previously presented in the manuscript. We now report the weighted LME models in the revised SI (**Tables S3B** and **S4B**).

Lastly, we decided against performing an analysis that penalizes the inclusion of more spatially distant electrodes in a cluster. As you note yourself, such an analysis would bias the results toward a more spatially segregated outcome, whereas we prefer to ask this question in a bottom-up, unbiased way. You said: "it would be useful to know if [the outcome of such an analysis] matches, e.g., existing patterns from fMRI, which are being challenged here". To clarify: we do not believe that our results directly contradict or challenge any existing fMRI findings (if you are aware of such findings, please let us know); instead, our results *extend* the fMRI literature by providing evidence of **distributed functional heterogeneity** within language cortex, which to-date has been argued to be largely functionally *homogeneous* (at least for the core, left hemisphere fronto-temporal network; e.g., Fedorenko et al., 2020 *Cognition;* Regev et al., 2021 *bioRxiv*).

3. Binary classifier. The authors place much weight on the fact that activation in Cluster 1 is able to discriminate S from W in the earliest time bins, while this takes 2-3 words to build up in Cluster 2. This is taken to reflect the time it takes Cluster 2 electrodes to "rule out" the existence of phrasal structure in W. However, unless I'm interpreting Fig. 5A incorrectly, it seems that Cluster 2 is reliably able to discriminate S from W in less than 500 ms, roughly the time it takes to present a single word in the fast-timing paradigm (450 ms). More generally, I assume "reliable discrimination" is taken to mean the first time point at which 10-fold cross-validated classifier accuracy differs significantly from chance (50%, one-sample t-test, reflected as stars in Fig. 5A). This analysis is likely too liberal because (i) assumptions of the t-test are often violated when applied to classifier accuracy (e.g., non-Gaussianity); (ii) the analysis does not account for temporal autocorrelation in the underlying time courses and interdependence between time bins introduced by overlapping windows (100 ms, 50% overlap); and (iii) there was no correction for multiple comparisons across time bins. In this sense, the classifier is very much a 1-D analog of MVPA in fMRI. Stelzer et al. (2013, Neuroimage) address the problems related to group-level inference on cross-validated classification accuracies. Short of their recommended approach, I would recommend at least adopting non-parametric approach estimation of the null distribution of group-level accuracy combined with multiple comparison correction across bins. The broader point though, is that inferences drawn with respect to when activation between conditions begins to differ are highly dependent on the statistical criterion. Moreover, little attention is paid to other aspects of the data, e.g., that classifier accuracy is much higher overall in the W vs. N contrast for Cluster 2 than Cluster 1. Finally, the J vs. N contrast (currently omitted) could be highly relevant here, given above comments related to J re: Cluster 2.

Thank you for these constructive suggestions – we agree that the binary classifier was not implemented in a sufficiently statistically rigorous way, and we have redesigned the decoding procedure following your suggestions (as described in the revised Methods section). Briefly: as before, we average the neural signal across 100 ms bins, but we have now moved to a 100 ms sliding window, which ensures that

adjacent timepoints do not contain any signal overlap. Furthermore, as you suggest, we have re-computed statistical significance using cluster permutation testing to control for the temporal autocorrelation structure in the signals and for multiple comparisons across time bins (Stelzer et al., 2013, *Neuroimage*; Maris & Oostenval, 2007, *J Neurosci*; see Methods for details).

The results from this new decoding procedure are shown in **Figure R6** below (and **Figure 3D-F** in the manuscript). We find that the Sentence and Word-list conditions can be reliably discriminated by Cluster 1 electrodes as early as the onset of the second word, whereas Cluster 2 electrodes can only discriminate the Sentences and Word-list conditions after the fifth word, and Cluster 3 electrodes can only discriminate these two conditions after nearly the entire timecourse is included in training. Cluster 1 electrodes also afford the best overall decoding performance when discriminating the Sentence and Word-list conditions. On the other hand, the Word-list and Nonword-list conditions can be discriminated earliest (around the onset of the third word) by Cluster 2 electrodes, whereas Cluster 1 electrodes cannot discriminate these two conditions until closer to the onset of the fourth word. The peak decoding performance is also much higher for Cluster 2 electrodes for this contrast. Finally, following your request, we have now included the Jabberwocky versus Nonword-list condition contrast. There, we see that—similar to what we find for the Word-list vs. Nonword-list conditions—Cluster 2 electrodes best discriminate these two conditions, showing both earliest discriminability and highest overall decoding performance.

As you mentioned, the time of initial discrimination depends on the sensitivity of the statistical method used; as a result, we no longer make strong claims based on these absolute decodability onset times and instead focus on the between-cluster differences. Thus, although the exact onset of discriminability between the Sentence and Word-list conditions is delayed (across clusters) under this new, more conservative procedure, the key patterns we had reported in the original manuscript hold. In particular, neural populations with a long TRW (Cluster 1 electrodes) are more strongly modulated by information that spans multiple words (which is present in sentences but not in word-lists), which manifests as earlier and overall stronger discriminability between the sentence and word-list conditions. In contrast, neural populations with a short TRW (Cluster 3 electrodes) are not strongly sensitive to information spanning multiple words, which manifests as poor discriminability between the Sentence and Word-list conditions. We have clarified these points in the Results section.

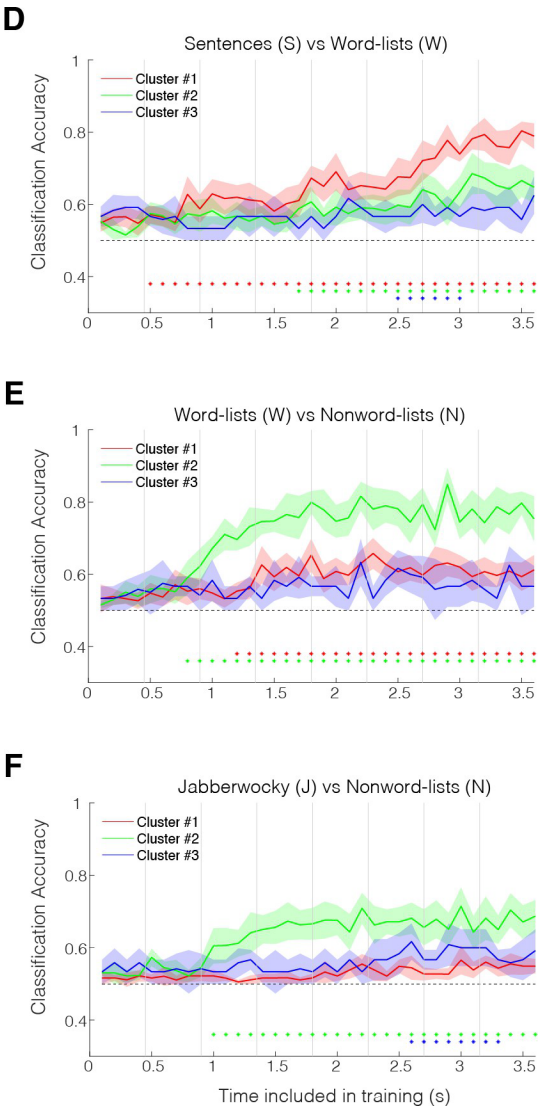


Figure R6 (panels D-F from Figure 3 in the revised manuscript). Classifier performance by cluster as a function of the amount of timecourse included in training (Methods). A binary linear classifier was trained to discriminate Sentence (S) and Word-list (W) conditions (**D**), Word-list (W) and Nonword-list (N) conditions (**E**), and Jabberwocky (J) and Nonword-list (N) conditions (**F**). Significance stars at the bottom (colored by cluster) reflect discriminability of conditions above chance level (ps<0.05, evaluated as a cluster statistic against a null distribution from permuted trials, Methods). Shaded regions reflect standard error across the 10 folds of the cross-validated classifier.

4. LME models. All three LME models (stimulus locking, TRW, and MNI coordinates across electrodes and participants) include a random effect of cluster within participant, where cluster membership is dummy coded as a categorical variable. It is unclear to me how the random effect of cluster was estimated given that all levels of cluster (1, 2, 3) were not present within all levels of participant. That is, Cluster 3 electrodes were present in only 4/6 participants (p. 7, but visually appears to be 5/6 in Figure 6B), which would result in a vector of all zeros for some

participants in the dummy coded variable column. Were the participants without Cluster 3 excluded from the LMEs? If not, is there evidence that the estimates of participant-level variance were robust and valid? The denominator DFs in the reported LME tables also seem high. If no corrective approximation (e.g., Satterthwaite, available in MATLAB fitlme) was applied during the computation of the ANOVA, justification is required as this could severely influence the threshold for significance.

Thank you very much for noting this issue. After the initial submission, we had discovered a small error in one of our visualization scripts that led us to mistakenly state that Cluster 3 electrodes were only present in 4/6 participants, when they were, in fact, present in ***all*** participants (see also our response below to Reviewer #2 regarding average cluster responses for individual participants, **Figure R18** and **Figure S1** in the revised manuscript). However, thanks to your comment, we have now re-evaluated the statistical rigor of our LME models and decided to re-run all the LME models with the *Satterthwaite corrective approximation method*, which is appropriate for relatively small sample sizes (Luke, 2017 *Beh Res Methods*). We have also changed the fitting method of the LME models to be REML rather than the default ML used previously, as this fitting method was shown to be most appropriate when using the Satterthwaite approximation (Luke, 2017 *Beh Res Methods*).

The use of this new, more statistically rigorous approach did not change the main results reported in the original manuscript. In general, the Satterthwaite method effectively decreases the degrees of freedom, which increases the p-values. However, most of the results remain significant, although a few of the more subtle comparisons now fall short of significance. Below we provide a summary of how the results changed with this new statistical modeling approach.

1. Comparing the degree of locking to individual word onsets, across clusters and conditions (**Table S1** in the manuscript)—only tested on Dataset 1:

Previously (prior to applying the Satterthwaite correction), all 3 between-cluster and 3 (out of 6) between-condition comparisons were significant, whereas now only 1 between-cluster comparison (clusters 3 vs. 1) and only one between-condition comparison (S vs. N) reach the alpha=0.05 significance level. However, an ANOVA for LME models revealed that the overall effect of cluster is significant whereas the overall effect of condition is not, similar to the original analysis.

2. Anatomical effects—comparing MNI coordinates across clusters (**Tables S2-4** in the manuscript):

➔ Dataset 1: Previously, out of 18 comparisons (3 between-cluster comparisons x 2 MNI coordinates x all/frontal/temporal electrodes) 5 were significant, whereas now only 4 are. The comparison that was previously significant and is not anymore is the z-coordinate comparison of Cluster 2 vs. 1 when examining temporal lobe electrodes.

➔ Dataset 2: Previously, out of 5 comparisons (1 between-cluster comparisons, (considering only clusters 1 and 3), x 2 coordinates x all/frontal/temporal electrodes) only one was significant (the z- coordinate comparison of Cluster 3 vs. 1 when examining all electrodes), whereas now no comparison reaches significance.

3. Comparing TRW estimates across clusters (**Tables S5-9** in manuscript):

➔ Dataset 1:

-All 6 participants included: all 3 between-cluster comparisons remain significant.

-Participants with a slow presentation rate (n=3): Previously, all 3 between-cluster comparisons were significant, whereas now the Cluster 3 vs. 1 comparison and the Cluster 3 vs. cluster 2 comparison fall just below the significance threshold (p=0.06).

-Participants with a fast presentation rate (n=3): Previously, all 3 between-cluster comparisons were significant, whereas now the Cluster 2 vs. 1 comparison is not significant.

(Note that the analyses that include subsets of participants with a slow or a fast presentation rate have only 3 participants and after the Satterthwaite approximation have very low degrees-of-freedom = ~2, resulting in a very low statistical power. Furthermore, this low number of degrees-of-freedom does not allow to run the ANOVA for LME, which we would need to run in order to test for the significance of the main effect of cluster, beyond comparisons of pairs of clusters.)

➔ Dataset 2: Unchanged (2 out of 3 comparisons remain significant).

Taken together, these issues primarily raise concerns related to (i) whether the three clusters identified correspond to three unique mechanisms primarily in terms of TRWs in the language system, or whether some other aspect of processing related to experimental condition plays a crucial role; and (ii) whether claims about distributed processing, derived from the spatial distribution of cluster-labeled electrodes, are supported robustly by the data.

Thank you again for all these helpful suggestions, which have helped increase the statistical rigor of our analyses. We hope that the revised analyses, along with the additional analyses we performed, help alleviate your concerns.

Minutiae:

I would like to see the behavioral results reported for both datasets. I found the results for four participants from Dataset 1 in Fedorenko et al. (2016), but full reporting for Dataset 1 and Dataset 2 would be welcome.

Yes, thank you. We have now included the behavioral results in the revised Methods section.

p. 24: “The following electrodes (electrodes; we use these terms interchangeably)”
p. 25, first line: A couple details should be added on how time warping was performed
p. 28, first line of “Electrode discrimination between conditions” section: examined examine
p. 28: “amplitude and offset of the sinusoidal by searching parameter combinations.” Sinusoidal should be either “sinusoidal function” or “sinusoid”
p. 30: “see 1 above,” I can figure out what this is referring to but not sure what ‘1’ means

All fixed, thank you.

Reviewer #2:

[summary omitted]

Overall the paper is clear and well written. I have some concerns that I outline in detail below. The most major of which relate to the specificity of the claims and some of the analytical approaches.

Thank you for the positive evaluation of this work and for all the helpful comments below.

MAJOR

[1] Vague interpretation

One difficulty I am having is understanding the broader implication of this result, and how it informs neural language processing. I think this is partly because the **differences across stimuli are fully ignored**, which makes it very difficult to link the difference in physiological responses across electrodes to the actual computation that is implemented. The authors do mention this shortcoming of analyzing responses "regardless of their structure or meaning", but their suggestion for overcoming this is to use single neuron recordings and deep learning language models. The data being analyzed here are some of the most powerful and valuable that the human neuroscience community currently has access to, and I believe that it is possible to address this shortcoming with the current dataset.

Specifically, how different are responses across trials but within a given condition? What explains that variance? Are all words processed the same, or is there variability? Does the "build-up" in activity have the same gradient across sentences? Is the gradual build-up actually there on all sentences or is it more jagged on an individual trial basis? To me, these are all important questions that would need to be answered in order to fully interpret the results. The clustering approach is useful for identifying different types of electrode dynamics but that seems like just the first step to actually understanding what those dynamics mean. The temporal resolution of the intercranial data are not being taken full advantage of, due to the averaging across different stimuli.

Thank you for highlighting the issue of inter-trial variability. As we have now emphasized more clearly in the revised manuscript, we whole-heartedly agree that both ***characterizing*** trial-level variability within a given condition, and ***explaining*** this variability by analyzing linguistic stimulus properties, is important for a complete neural account of language processing. Our response to this comment is in three parts: 1) we briefly defend the feasibility of investigating language processing mechanisms, including informing linguistic computations, by examining ***condition-level differences***; 2) we show that our main findings are ***robust to inter-item variability***; and 3) we discuss the reasons, and provide some evidence, for why the

current data are, unfortunately, ***not well-suited for investigating stimulus-related responses*** beyond condition-level differences.

1) Condition-level differences have always been critical in investigations of language processing mechanisms.

You wrote: "*One difficulty I am having is understanding the broader implication of this result, and how it informs neural language processing. I think this is partly because the differences across stimuli are fully ignored, which makes it very difficult to link the difference in physiological responses […] to the actual computation that is implemented.*"

Although we agree with you that examination of stimulus-level properties can provide important insights, it is important to remember that language research has long benefited from *comparisons of classes of stimuli*: from examination of naming latencies for pictures with high vs. low frequency names (e.g., Oldfield & Wingfield, 1965 *QJEP*; Jescheniak & Levelt, 1994 *JEP:LMC*), to measuring reading times for syntactically complex vs. simpler sentences (e.g., Just & Carpenter, 1992 *Psych Rev*; Grodner & Gibson, *2005 Cog Sci*), to examining event-related potentials in EEG data to sentences that end in more vs. less likely words (e.g., Kutas & Hilliyard, 1980 *Science*) or to sentences that do vs. do not contain grammatical violations (e.g., Hagoort et al., 1993 *LCP*). In fact, arguably, i) most key discoveries in language research were made based on *condition-level differences*, and ii) most theories and computational models of both lexical access and syntactic structure building were developed and evaluated largely using data from *condition-level manipulations* (cf. examination of individual-item-level responses). Of course, ever since Herb Clark's 1973 seminal paper on the importance of accounting for inter-item variability in language research, a lot more attention has been paid to properly handling item effects in statistical analyses of condition-level differences. However, condition-level effects have always played—and continue to play—a critical role in advancing theories of language processing.

2) Robustness of condition-level effects to inter-item variability.

Before evaluating inter-item variability, it is first important to establish that our observed condition-level effects are robust across items (i.e., hold for most items), rather than being driven by a small number of individual items (e.g., you asked: *"Does the "build-up" in activity have the same gradient across sentences? Is the gradual build-up actually there on all sentences or is it more jagged on an individual trial basis?"*). In the original manuscript, we already showed that our main result of the three response profiles is robust to removing half of the trials (**Figure 3A**). However, this is a relatively coarse-level approach. To evaluate how similar item-level responses are to the mean effects we report, we performed an additional analysis where we correlated randomly sampled item-level timecourses (i.e., concatenating a random S, W, J and N trial timecourse) with the average timecourse (over all items). We performed this procedure separately for each cluster, and a given item timecourse was averaged over the electrodes in each cluster. This analysis effectively asks: how similar is any given 'item' (i.e., a set of an S, a W, a J, and an N item) is to the condition average. We repeated this procedure 100 times for each cluster and we found that the average correlations (across the 100 repetitions) for all clusters were

positive and in the moderate-to-high range (0.47 for Clusters 1 and 2 and 0.33 for Cluster 3). Thus, although in our main analyses we average timecourses across items, this analysis establishes that the individual-item timecourses are representative of the cluster's average profile, so the findings are robust to inter-item variability.

For completeness, we performed a similar analysis for each condition separately, because the correlations in the analysis just presented are driven, to some extent, by the between-condition differences. Even for each condition separately, the correlations are quite strong—especially for the Sentence condition which is the focus of our TRW investigations (0.31-0.55 for the Sentence condition across the three clusters; 0.17-0.38 for the Word-list condition; 0.16-0.25 for the Jabberwocky condition; and 0.14-0.33 for the Nonword-list condition, **Figure R7B**). Furthermore, most individual items are positively correlated with the average, and the distributions appear approximately normal (cf. a bimodal distribution which would indicate that there exist distinct types of items in our stimulus set, **Figure R7B**).

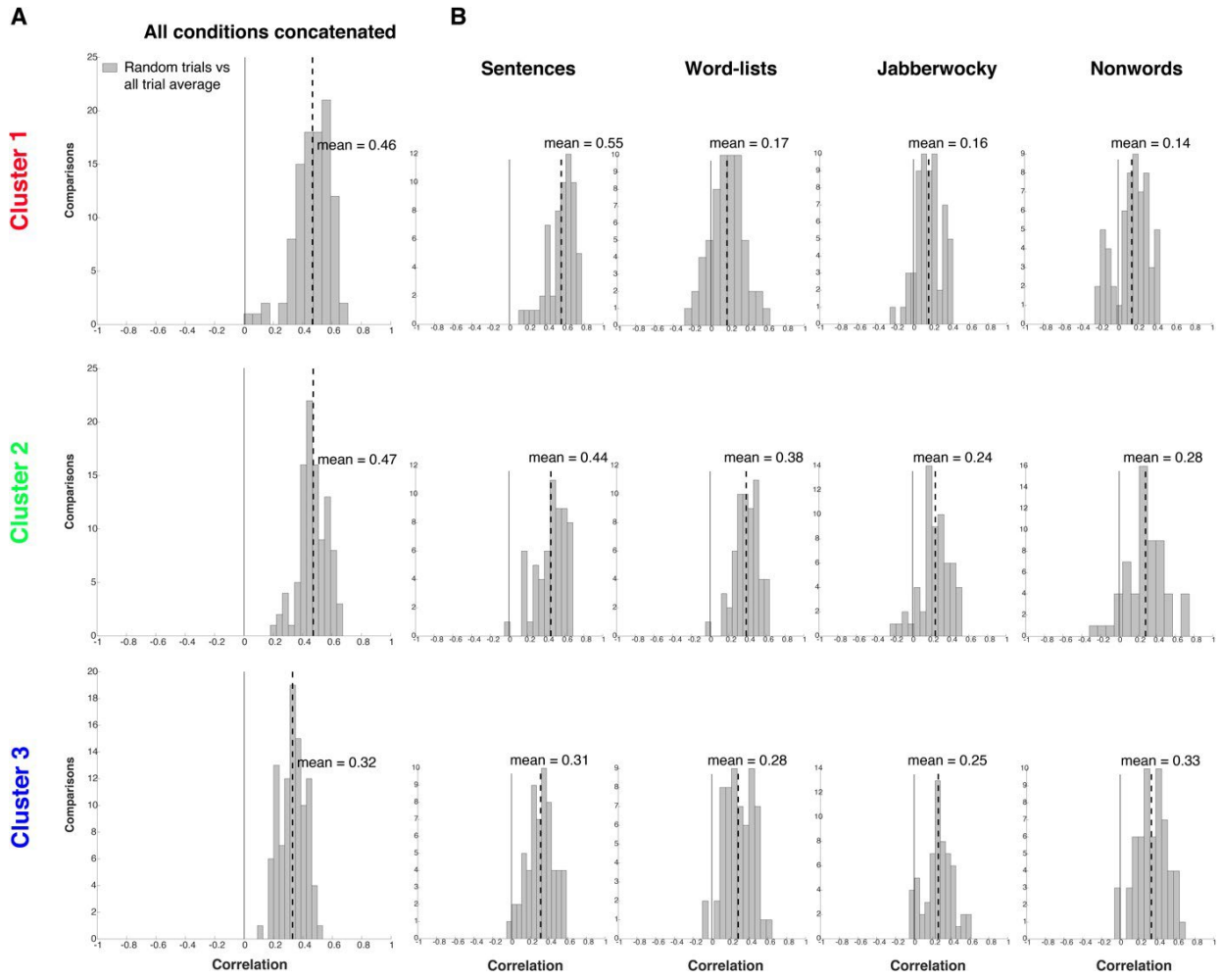


Figure R7. Correlations between single trials and the average. Correlations were computed for each cluster separately, averaging across all electrodes per cluster. Correlations were computed for all conditions concatenated (as was done for our clustering analyses, **A**), or for just the Sentence, Word-list, Jabberwocky or

Nonword-list conditions (**B**). Histograms are of the correlations between 100 randomly selected trials and the average.

3) The datasets in this study are not well-suited for investigating item-related responses beyond condition-level differences.

We spent quite some time—when working on the original manuscript, and some more since getting the reviews back—thinking about investigating item-related responses beyond condition-level differences, so please exclude our lengthy response to this point. Unfortunately, as elaborated below, for a combination of empirical, methodological, statistical, and conceptual reasons, we have concluded that the current datasets are not well-suited for such an investigation.

a) No *a priori* hypotheses about the modulation of neural activity by lexical or syntactic properties.

Given our interpretation of the cluster differences in terms of their TRWs, it is certainly possible to generate some hypotheses about how different lexical (word-level) and syntactic variables might affect activity in electrodes that belong to the different clusters. However, such hypotheses would be *a posteriori*, and evaluating them with a requisite level of rigor would almost certainly require additional data collection on new experimental materials that are designed to maximize variance along the hypothesized dimensions of interest (see also pt. d below for evidence of the lack of sufficient variance in the current materials). Without *a priori* (ideally, pre-registered) hypotheses, we are hesitant to engage in a bottom-up exploration of potential lexical/syntactic effects as the set of potential analyses is ~unbounded. Intracranial data are rich and multi-dimensional, which makes it easy to find "significant" effects in some electrode(s) for some variables of interest, but we would not trust such effects. Given the current emphasis in the field on robustness and replicability, we prefer to focus on our core finding of three distinct temporal profiles. We agree with you that this finding is just "the first step in understanding what [these profiles] mean", but this doesn't undermine the finding's importance.

Moreover, because the finding replicates in an independent dataset, we have confidence in it, and we— as well as, we hope, others—will build on it in future work. The field is chock full of spurious findings; we do not want to add to this state of affairs.

b) No reliable stimulus-related activity detected at the cluster level (averaging across electrodes within each cluster).

In order to examine the effects of linguistic features on neural responses to individual items, which are inherently noisy, it is critical to first establish that the relevant neural responses are *sufficiently modulated by stimulus properties*. That is to say, stimulus-related response variance must exceed the variance attributable to measurement noise. To test for this, one can compare the similarity of neural responses to the same item versus the similarity of responses to different items. Ideally, this within- vs. between-item comparison should be performed within the same participant and electrode. However, because in the current study any given stimulus was only presented once to each participant, we performed this analysis *between participants,* which, in turn, required averaging across electrodes within each cluster (to establish between-participant correspondence). In particular, for each

participant, we averaged the timecourses for each item across the electrodes within each of the 3 clusters, which resulted in 3 timecourses per participant per item. We considered the items in the Sentence (S) and Word-list (W) conditions only, which contain real words, as would be required for examining modulation by lexical variables. Then, within each cluster, we computed correlations for each pair of participants for each item (asking e.g., how similar is the timecourse for the sentence "*Dad was tired so he took a nap*" in the Cluster 1 electrodes of Participant 1 vs. in the Cluster 1 electrodes of Participant 2). We then computed correlations for each pair of participants between the timecourse for a given item and the timecourse for a different item (e.g., "*Beth walked her dog in the park nearby*").

The distribution of these within-item and between-item correlations for the three clusters for the Sentence and Word-list conditions are shown in **Figure R8** below. The means of the distributions of both the within- and between-item correlations are positive but close to zero and not different from each other, as would be required to examine response modulation by stimulus properties. So, although the condition-level effects are robust to inter-item variability (pt. 2 above), we *do not see evidence of similar timecourses for the same item across participants*. This result suggests that i) stimulus repetitions within a participant may be required to obtain reliable item-level responses, and/or ii) greater linguistic variability is necessary in order to differentiate item-level responses (see pt. c below). Another possibility is that iii) by averaging responses across the electrodes within a cluster we lose relevant information (which could be the case if only a subset of electrodes in a given cluster reliably responded to stimulus features, or if different electrodes in a cluster responded to different linguistic features).

However, as noted above, without electrode averaging, establishing between-participant correspondence is not possible, which makes it challenging to ensure that we are not modeling noise (as we have no within-participant stimulus repetitions).

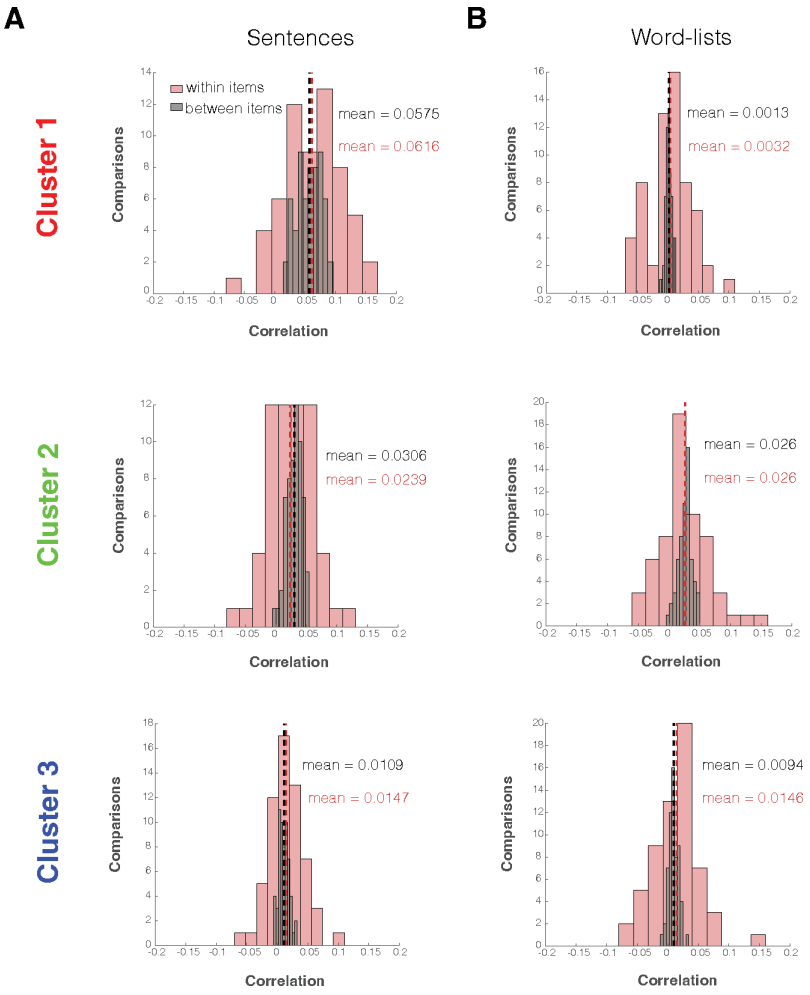


Figure R8. Correlation of individual item timecourses across participants within the same item (red) and across different items of the same condition (black) for the sentence (**A**) and world-list (**B**) conditions. Item timecourses are constructed by averaging across all electrodes in a particular cluster within a given participant.

c) The experimental language materials are not well-suited for investigating the modulation of neural responses by lexical or syntactic features of the stimuli.

Examining sensitivity to different linguistic properties, e.g. of *particular words* and/or *particular syntactic structures*, requires both sufficient variability within the stimulus set in those linguistic properties and an ability to disentangle various properties from one another. Due to practical, as well as scientific, decisions that were made at the time of material creation for this experiment (briefly expanded on below), neither of these conditions are met in the current stimulus set. First, we wanted to make this experiment *feasible for diverse populations*—including relatively lower-functioning patients with epilepsy (as relevant to the current study), non-native English speakers, and children. The sentences are short–which limits the kind of syntactic complexity that can be present–and use common, easy-to- process syntactic structures, as well as frequent and relatively short words. Additionally, as the goal of

the experiment was to examine condition-level differences, we did not worry about dissociating different linguistic features that typically co-vary in natural language.

The lack of sufficient variability in lexical and syntactic properties and the co-variation among linguistic features in the current materials is illustrated in **Figures R9** and **R10**. For example, in **Figure R9**, we plot tree depth, a measure of the number of incomplete syntactic dependencies that must be maintained in memory at any given point in the sentence (e.g., Gibson, 1998 *Cognition*). As you can see in **Figure R9A**, tree depth scales with word position in our sentence stimuli, making it a *descriptive feature* that could potentially explain the observed buildup of neural activity in Cluster 1 electrodes. However, as you can see in **Figure R9B, C**, the tree depth values for the vast majority of the 80 sentence trajectories in Dataset 1 generally increase over the course of the sentence. Testing whether this feature modulates the build-up of neural activity would require sentences with more diverse structures (e.g., left-branching structures, where tree depth is high earlier in the sentence, decreasing in the later positions). The sentences would also need to be longer in order to allow for greater syntactic diversity across the sentence.

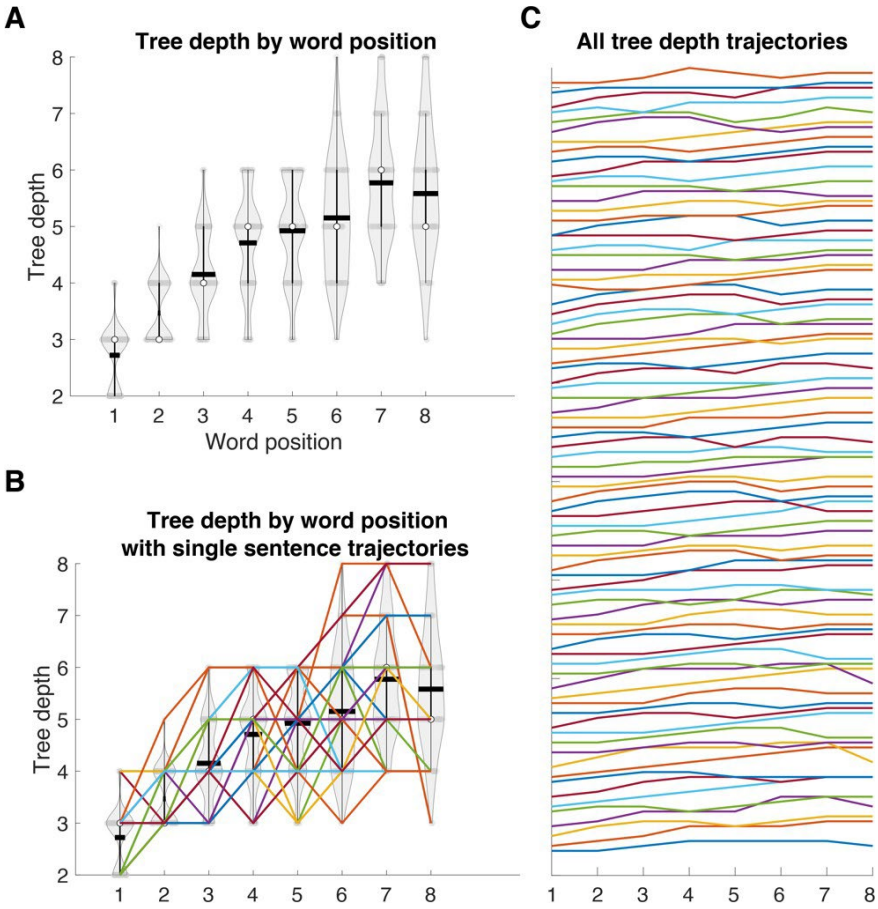


Figure R9. Syntactic tree depth computed from our sentence stimuli. All 80 sentence stimuli were first parsed into binary-branching phase-structure trees using the Berkeley parser (Petrov & Klein, 2007 *ACL*). Then, for each sequential word in the stimuli, tree depth was computed as the number of constituents above that particular word in the tree. **A**) For each out of the 8 word positions in the sentences, the distribution of tree depths across the 80 sentences is plotted as a violin plot (horizontal black line represents the mean, white circle the median, and gray

dots represent all individual values). **B**) On top of the violins we add individual sentence trajectories for each specific sentence item. **C**) The same trajectories as in B, but to facilitate viewing individual trajectories, here each trajectory is displaced upwards relative to the previous trajectory.

And in **Figure R10C**, we show that word length and word frequency show a standard Zipfian correlation in our materials (more frequent words tend to be shorter; Zipf, 1949; Piantadosi et al., 2011 *PNAS*), and that the ranges of both lengths and frequencies are quite restricted. In **Figure R10A, B**, we show that unigram word frequencies highly correlate with both bi-gram and tri-gram frequencies (i.e., how commonly a word occurs in the context of the preceding word or the preceding two words), which makes it difficult to disentangle lexical from contextual effects in these materials.

For all of these linguistic features, it is *possible* to construct materials, where they vary substantially, as well as dissociate from one another (and from features like word position within the sentence); it is just that the current stimulus set was not constructed with the goal of examining modulation by linguistic features at the item level.

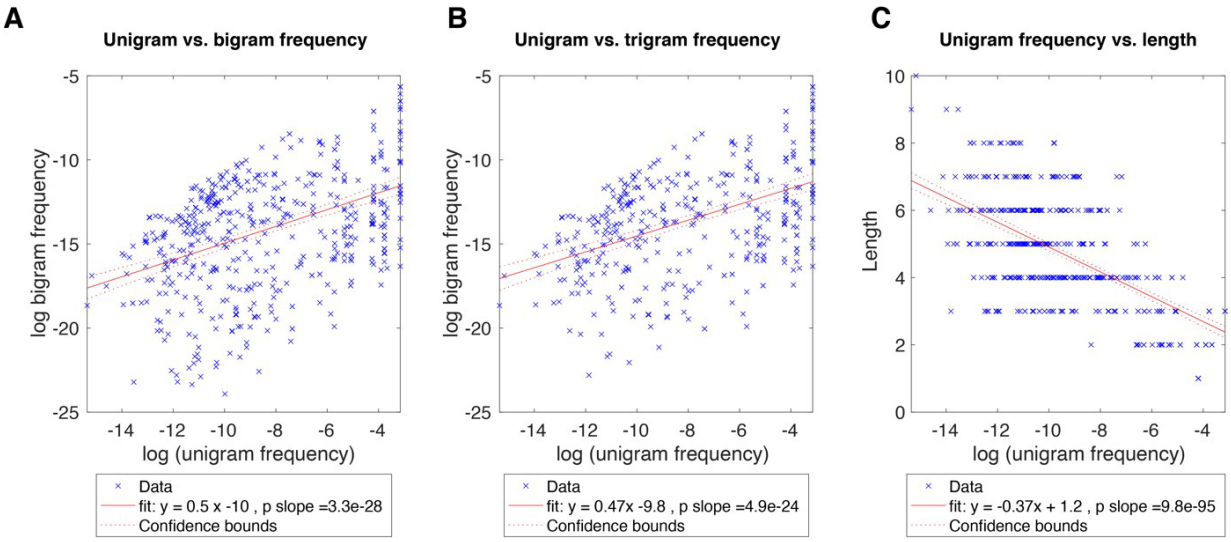


Figure R10. Correlations between linguistic features within our stimulus set. Linguistic features were computed for each individual word within the sentence stimuli. Unigram, bigram, and trigram frequencies were extracted from the Google Ngram online platform (https://books.google.com/ngrams/) using our own script, averaging across google books corpora between the years 2010 and 2020. Length was computed as the number of letters in a word. A linear fit was computed using a linear model, and is presented on top of the scatter plots in red. The intercept, slope and p-value of the slope are presented in legends. All correlations are large and significant.

This is all to say that even if we constructed some hypotheses *a posteriori* (concern a above), considered individual-electrode responses rather than cluster-level responses (concern b), and found that some linguistic variable(s) modulate activity in certain electrode(s) in this limited dataset, we would be unable to unambiguously attribute the effects to a particular linguistic variable, as it would be correlated with other variables.

For all these reasons (which we now succinctly summarize in the manuscript), we prefer to not explore stimulus-related responses; instead, in the Discussion we outline a few *hypotheses* about the sensitivity of the electrodes in the three clusters to different lexical and syntactic variables, which can be pursued in future work. In addition, the individual item timecourses will be made available on OSF upon publication, which would allow other researchers to perform exploratory linguistic analyses.

[2] Modeling approach

In a similar vein, I had some concerns regarding the modeling. I could not find a description of what hypothesized neural generators were being modeled here, and what the underlying hypothesis is of what processes are actually happening. This felt more like a way of capturing the dynamics of the already observed time courses rather than a prior conceived hypothesis. That is fine, but it needs to be discussed as such. And, framed as an analysis method not a model, if that is the goal.

Building on that, the authors use a gaussian kernel with variable widths to estimate the different window sizes. These kernels are centered around the onset of each word, by virtue of the convolutional method. This means that the neural activity is assumed to increase prior to the word being presented on the screen, and the increase is assumed to be symmetric with the decrease. Why are these assumptions built into the model? In a few places, the authors refer to these processes as oscillatory. If that is their mechanistic assumption, then the process should be modeled as such, with sinusoids rather than cumulative gaussian kernels. Otherwise, it should be modeled as a series of evoked responses with a **neurophysiologically realistic response function**, and a hypothesis for the underlying mechanistic generator. In the most ideal case, this would be performed on, and be able to account for variance in, the single trial data.

Thank you for the chance to clarify. You are absolutely right that the modeling is really a way to **capture the observed condition-level dynamics**, not an *a priori* hypothesis; our goal was to illustrate how a simple instantiation of an information processing system, with one (interpretable) free parameter (the length of past stimulus context), could give rise to the response dynamics that we observe. We chose to follow the field's convention of referring to such "toy models" as a model (e.g., Chang et al., 2022 *PNAS*), but we agree that the description and the construal of this approach in the original manuscript were not clear. We have now revised the text in the relevant section to explain the approach more clearly and we now refer to the model as a "toy" model.

Below we clarify a few more specific points relevant to your comment.

1) Our use of a convolution does not imply an oscillatory neural generator but rather an evoked response.

Although the underlying mechanistic generator of the observed neural signals is a matter for future research, we ***do not see the process as oscillatory*** (and we fixed our wording in the manuscript to reflect

that) but rather as a series of evoked responses triggered by each word presentation, which sounds like it aligns with your views. Furthermore, regarding your concern that *"the neural activity is assumed to increase prior to the word being presented on the screen"* in our analyses: using a convolution between a stimulus train and a kernel does not imply an activity increase prior to stimulus presentation. In fact, the way that we implement the convolution is equivalent to assuming that the kernel function is the neural response evoked by the presentation of each new stimulus, and this response kernel appears right *after* each stimulus presentation. To illustrate this point, we repeated the temporal receptive window (TRW) estimation, with cosine kernels, twice - one time using a convolution, as we did in the manuscript, and the other, simply replacing each word onset with the kernel function right *after* it, to produce the simulated electrode response. We show below that these two methods are equivalent - the predicted electrode responses (**Figure R11 A, B**) as well as the final TRW estimations for all electrodes are identical using these two methods. We now also clarify this point in the manuscript. 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.

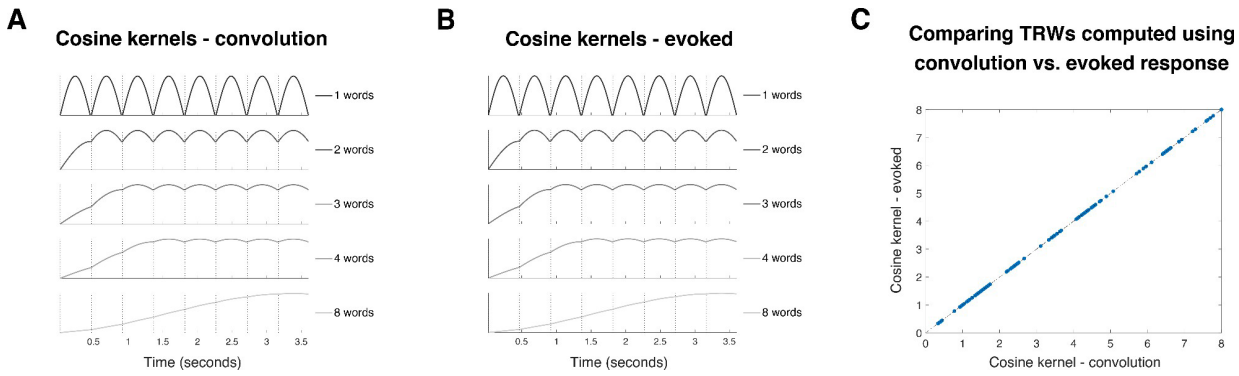


Figure R11. Convolving kernels with a stimulus train is identical to assuming evoked responses in the shape of the kernel. We fitted the temporal receptive window (TRW) model using Cosine kernels two times; First, as reported in the paper, using a convolution between a simplified stimulus train (giving a value of 1 at stimulus onset and zero elsewhere) and the kernel function, and second, by placing an evoked response at the shape of the kernel beginning at each stimulus onset. These two methods are identical. **A-B)** Simulated signals for TRW sizes of 1,2,3,4, and 6 words, are identical across the two methods. Vertical lines represent the onsets times of each of 8 word stimuli (including the y-axis itself which represents the onset of the first word). **C**) Each dot in the scatter plot represents an electrode, x-axis is the best fit TRW using the convolution method and y-axis is the best fit TRW using the evoked method. The two values are identical for all electrodes.

2) Simplifying assumptions of our toy model

We acknowledge that the assumptions built into our toy model should have been explicitly stated. We expand on three key assumptions below and mention them in the revised manuscript. ***First***, we assume that the minimal stimulus unit of interest for language-responsive neural populations is a **word**, as reflected in our choice of a stimulus train that corresponds to *words* rather than sub-lexical units or multi-word sequences. Although this is a simplifying assumption (e.g., because the language network is sensitive to sub-lexical structure; Bozic et al., 2010 *PNAS;* Regev et al., 2021 *bioRxiv*), words appear on

the screen one at a time in our experimental paradigm, so we assume that at each stimulus presentation, word-related processing takes place. Furthermore, we ignore potential differences between the words as we model each word as a simple delta function at its onset (see our reply to your previous question for a discussion about why we do not find this dataset suitable for modeling linguistic differences between stimuli).

Second, in light of the observed dynamics, we assume that each population has a ***fixed*** temporal receptive window over which it can process information. This may also be a simplifying assumption–the receptive window of a neural population may be adaptable to particular linguistic context, e.g., TRWs might lengthen in environments that require integrating longer spans of linguistic information for efficient understanding, like in complex or noisy linguistic environments. However, there is limited linguistic diversity in our materials, so we assume that even an adaptable TRW would remain more or less fixed in response to these stimuli. Future studies should investigate whether the receptive window of a given neural population adapt to linguistic context (Discussion).

And ***third,*** we make some assumptions about the shape of the evoked response (kernel) in our toy model (i.e., a smoothed box car function). Most importantly, we wanted the response function to capture the full assumed temporal receptive window (cf. long-tailed response functions that have a shorter "effective" receptive window) to make the interpretation of the fitted window size straightforward. In addition, we assumed that neural activity in response to a stimulus unit of interest would increase and decrease *gradually* over time (cf. an abrupt change of voltage such as in a boxcar shape). The reason for this assumption is that ECoG/sEEG electrodes sum up neural activity from a population of neurons and it is unlikely that the activity of all these neurons will rise or fall all at once. We acknowledge that these two minimal considerations will not deliver a neurophysiologically realistic response function; however, we attempted to make the fewest possible assumptions about the function's shape and we do not make a strong claim about the kernel function being representative of the true neural response. Rather, we take the toy model as a means to tap into the *relative* temporal receptive windows across electrodes. We have now explicitly stated these assumptions in the revised manuscript (Methods).

3) The relative difference in TRW by cluster does not depend on kernel shape.

To address your comment, we have also repeated the same TRW fitting procedure as in the original manuscript using five different response functions shown in **Figure R12** (and **Figure S6** in the revised manuscript) to establish that our key result—that the TRW size decreases from Cluster 1 to 2 to 3—does not strongly depend on the choice of response function. More specifically, we test cosine, "wide" Gaussian (as used in the main analysis), "narrow" Gaussian, square (i.e., boxcar), and linear asymmetric functions (**Figure R12A-E**). Importantly, ***our main result is robust to the choice of kernel shape.***

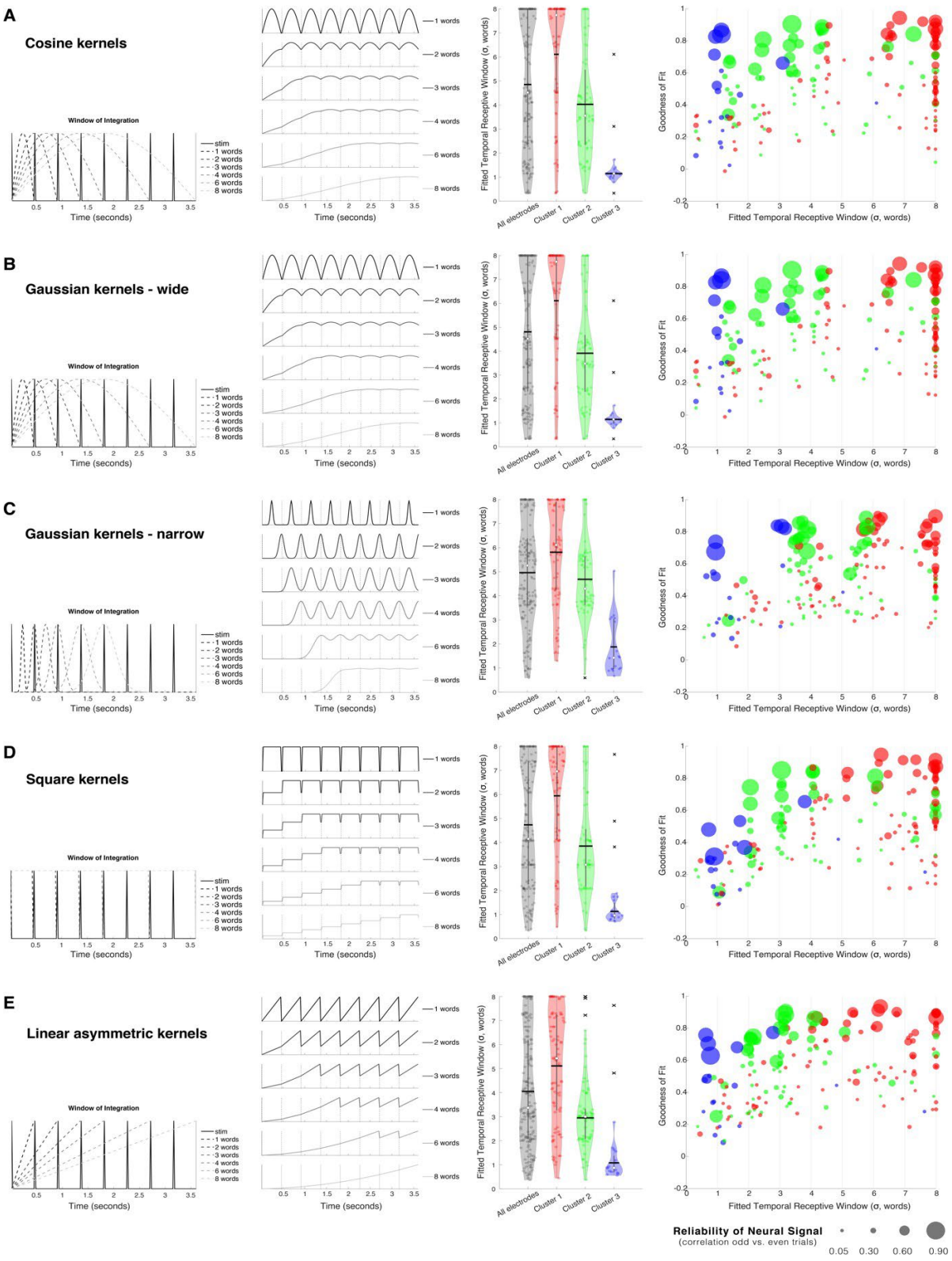


Figure R12 (Figure S6 in Supplementary Information). Temporal receptive window (TRW) estimates with kernels of different shapes. Modulation of TRW by cluster does not depend on the choice of kernel. TRW model was fitted

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 the original manuscript; **B**), "narrow" Gaussian (Gaussian curves with a standard deviation of $\sigma/16$ that were truncated at +/- 8 standard deviations; **C**), a square (i.e., 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 **Figure 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 and electrode) 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 show a decreasing trend from Cluster 1 to 3.

3) Our toy model can be extended to include condition-level linguistic features.

Finally, although our current toy model can only account for the responses to the Sentence condition, to move our approach one step closer toward explaining the full response profile, we now elaborate in the Discussion on a way in which this toy model could be extended, in principle, to also describe the responses across linguistic conditions. In particular, we propose that neural populations in the language network are not only sensitive to different lengths of past context, but that the neural response of these populations scales with how well the stimulus matches the statistics of natural language (effectively, how "language-like" the stimulus is) ***at the relevant context length***. For example, a neural population that integrates 2-word contexts would modulate its response based on the plausibility of the encountered bigram (words *n-1* and *n*, where n is the current word), whereas for a population integrating 6-word contexts, the response would be modulated by the plausibility of the associated 6- gram (word *n* and the preceding 5 words). Furthermore, we illustrate that sentences and word-lists increasingly diverge in their language-like-ness as context length grows in our set of stimuli (**Figure R13** below and **Figure S5** in the revised manuscript). Therefore, we propose that the dependence of the response on how language-like the input is—in combination with temporal receptive windows (TRW)— could explain why Cluster 1 populations, which track longer contexts, show a more pronounced difference in response to the Sentence and Word-list conditions, compared to Cluster 2 and 3.

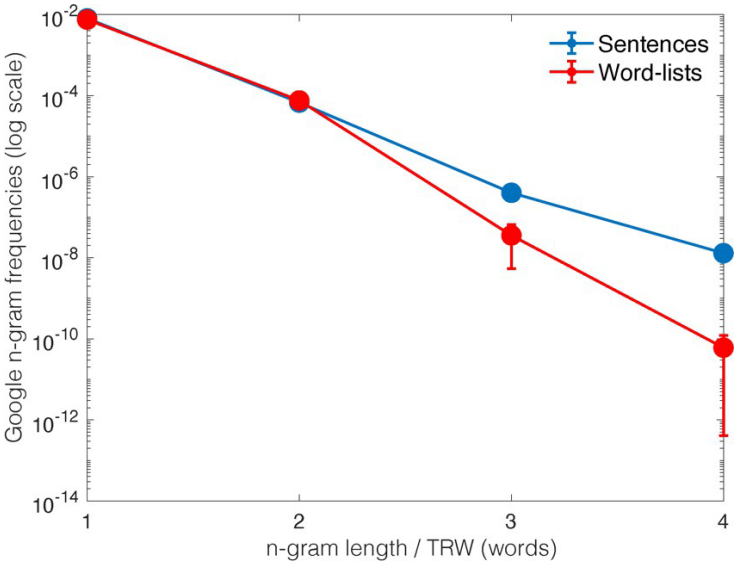


Figure R13 (Figure S5 in Supplementary Information). N-gram frequencies of sentences and word-lists diverge with n-gram length. N-gram frequencies were extracted from the Google n-gram online platform (https://books.google.com/ngrams/), averaging across Google books corpora between the years 2010 and 2020. For each individual word, n-gram frequency for n=1 is the frequency of that individual word in the corpus, for n=2 is the frequency of the bigram (sequence of 2 words) ending in, and including, that word, for n=3 is the frequency of the trigram (sequence of 3 words) ending in, and including, that word, etc. 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 word 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 wordlists grows as a function of N.

In sum, as we have now elaborated in the Discussion, we suggest that our toy model, which captures the "gating" of linguistic input into different lengths of effective input due to a characteristic TRW, can be extended to include a further component where the effective input's probability is evaluated (against past linguistic experience) and the neural responses are scaled proportionally. This proposal moves us closer to stimulus-dependent accounts of linguistic computations carried out by different neural populations in the language network and can be evaluated in future experiments by manipulating linguistic probability of n-grams at varying scales.

[3] Cluster-electrode assignment

The paper relies heavily on assigning electrodes to clusters. It would be useful to understand how "pure" an electrode is relative to its cluster assignment. Does it also have significant weight on the other clusters? How much of the other timecourse dynamics are intermixed within an electrode response? Given that the results suggest that the spatial organization is essentially random, it would seem likely that a single electrode would capture responses from neural populations that also contain different dynamics. Understanding to what extent this is the case would be helpful for better interpreting the heterogeneity of the results.

Thank you for these questions. We agree that it is certainly possible for a single electrode to have a response profile that falls somewhere between the prototypical response profiles (e.g., if there exists a gradient between response types or if an electrode is picking up activity from multiple neural populations). To evaluate the extent to which the observed responses can be attributed to a single profile, we computed partial correlations of every electrode's mean timecourse with that of each of the cluster medoids, while controlling for the correlations with the other two medoids (**Figure R14**). This analysis asks how much a particular electrode's correlation with a given cluster medoid can be uniquely attributed to that cluster, as opposed to the other clusters. If a neural population has a "mixed" response profile, one would expect its partial correlation with another cluster's medoid (when the correlation of its assigned cluster medoid is excluded) to be high, indicating that the population's response cannot be fully explained by either cluster alone. As shown in **Figure R14**, this is not the case for most electrodes, suggesting that many of the electrodes exhibit response profiles that are mostly consistent with *one* of the prototypical responses (i.e., relatively "pure").

Nonetheless, a few electrodes in both Clusters 1 and 2 (but not Cluster 3) exhibit high partial correlations with another cluster's medoid (i.e., a "mixed" response profile, **Figure R14B**). Visual inspection of these response profiles reveals a blend of Cluster 1 and Cluster 2 response characteristics (**Figure R14C, D**). The existence of mixture electrodes primarily between Clusters 1 and 2 is in line with the high correlation between their medoids (0.68 between Cluster 1 and 2 medoids versus 0.21 between Cluster 1 and 3, and 0.24 between Cluster 2 and 3; **Figure 3A** in the revised manuscript). These "mixed" populations could either indicate that a continuum of responses between the cluster medoids is, to some degree, present in neuronal responses, or it could be that some electrodes are capturing responses from multiple neural populations. In order to adjudicate between these two alternatives, a higher spatial resolution is needed. We now added this issue to the Discussion.

In addition, we plot the anatomical distributions of electrodes shaded by their partial correlation with each of the cluster medoids (**Figure R14E**). This visualization shows the anatomical trend of "Cluster-1- like" or "Cluster-2-like" responses, regardless of the electrodes' actual cluster assignments. This figure demonstrates that whereas Cluster-1-like and Cluster-2-like responses are still found throughout frontal and temporal regions (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)), Cluster-3-like responses are strongly localized to the posterior STG. We have included **Figure R14** as **Figure S4** in the revised manuscript.

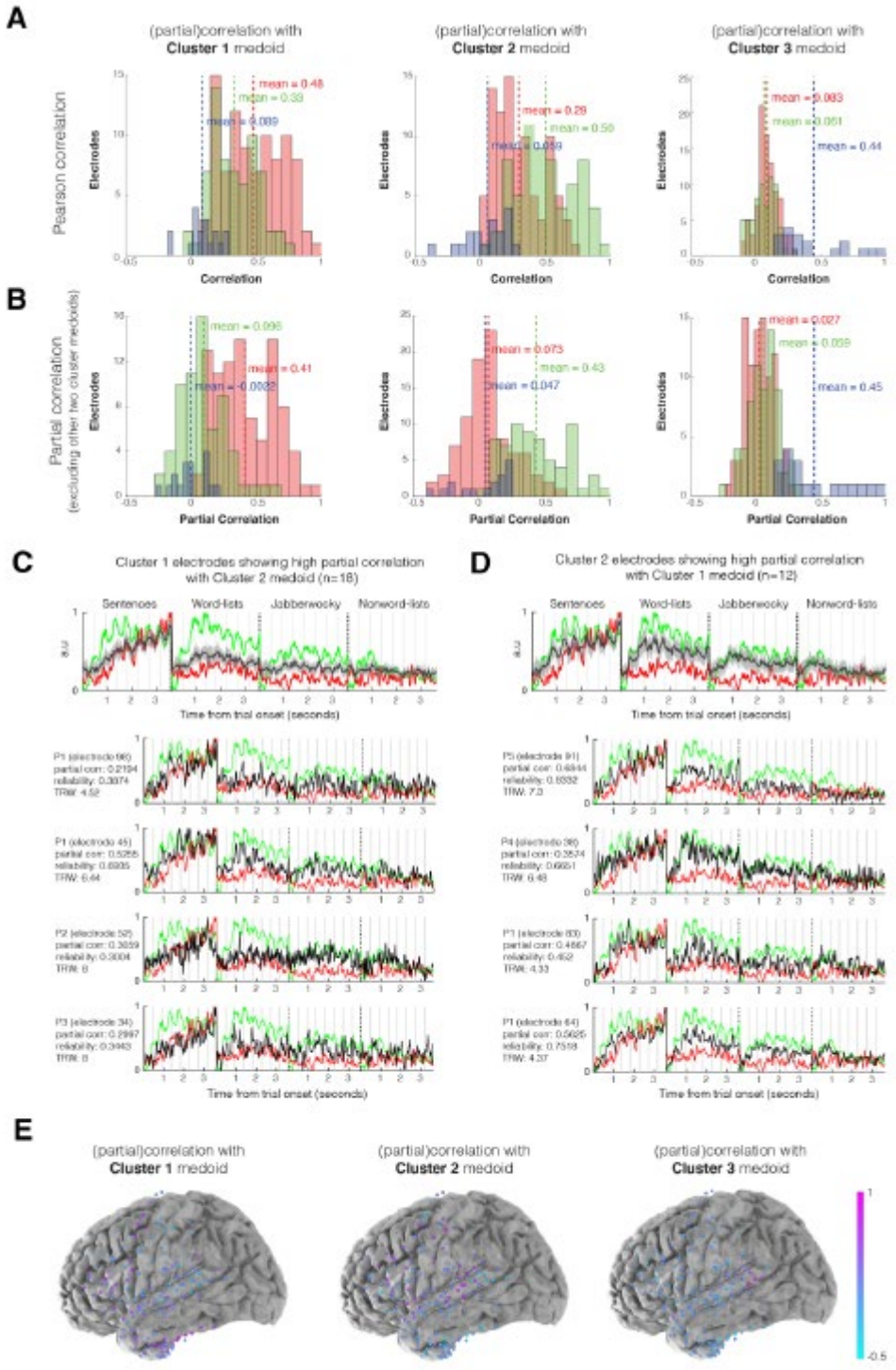

A
(partial)correlation with Cluster 1 medoid
(partial)correlation with Cluster 2 medoid
(partial)correlation with Cluster 3 medoid
Pearson correlation
Electrodes
Correlation
mean = 0.48
mean = 0.33
mean = 0.089
mean = 0.29
mean = 0.59
mean = 0.059
mean = 0.083
mean = 0.051
mean = 0.44
B
Partial correlation (excluding other two cluster medoids)
Partial Correlation
mean = 0.098
mean = 0.41
mean = -0.0032
mean = 0.073
mean = 0.43
mean = 0.047
mean = 0.027
mean = 0.059
mean = 0.45
C
Cluster 1 electrodes showing high partial correlation with Cluster 2 medoid (n=18)
Sentences
Word-lists
Jabberwocky
Nonword-lists
Time from trial onset (seconds)
D
Cluster 2 electrodes showing high partial correlation with Cluster 1 medoid (n=12)
E
(partial)correlation with Cluster 1 medoid
(partial)correlation with Cluster 2 medoid
(partial)correlation with Cluster 3 medoid

Figure R14 (Figure S4 in the Supplementary Information). Partial correlations of individual response profiles with each of the cluster medoids. **A)** Standard Pearson correlations of all response profiles with each of the cluster medoids, grouped by cluster assignment. **B)** Partial correlations of all response profiles with each of the cluster medoids, excluding the correlation of the response profile with the other two cluster medoids, grouped by cluster assignment. **C)** Response profiles from electrodes assigned to Cluster 1 that have a high partial correlation (>0.2, arbitrarily defined) with the Cluster 2 medoid (and split-half reliability>0.3). *Top:* Average over all electrodes that meet 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 have a high partial correlation (>0.2, arbitrarily defined) 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 correlation with a given cluster medoid (controlling for the other two medoids).

MINOR

* I found the motivation for using invasive recordings a little surprising, in terms of claiming that fMRI has poor spatial resolution. Certainly the temporal resolution is better for neurophysiological recordings, but the blanket statement that fMRI has poor spatial resolution was puzzling. What about high Tesla systems for example? In addition, the electrodes being used here are a couple of millimetres in diameter with a spacing of 1cm — this is not such excellent spatial resolution within a single subject. I would recommend that the authors remove entirely the claim about spatial resolution differences.

Thank you – we have adjusted the relevant portion of the introduction.

*The authors seem to be heavily citing their own work, and missing out important and relevant research from other research labs.

We have tried to add more references to other labs' work throughout the manuscript. If you see any glaring omissions, please let us know, and we would be happy to include them.

* The authors do a good job of demonstrating that the clusters they find are robust across exclusive partitions of the data. And I understand that the authors motivate the use of 3 clusters by use of the "elbow" method. I also appreciate adding different 'k' values in the supplementary. However, the amount of variance accounted for by each of the first 10 clusters may still be interpretable, beyond the fourth cluster. It would be useful to include visualization of each of the 10 clusters similar to that in Figure 3D.

Thank you – we agree and have now included a figure in SI (**Figure S2** and **Figure R15** below) that implements k-medoids clustering with k=10 for Dataset 1. This figure further demonstrates how dominant the thee reported response profiles are, as the responses from many of the additional clusters when clustering with k=10 resemble the prototypical responses when clustering with k=3.

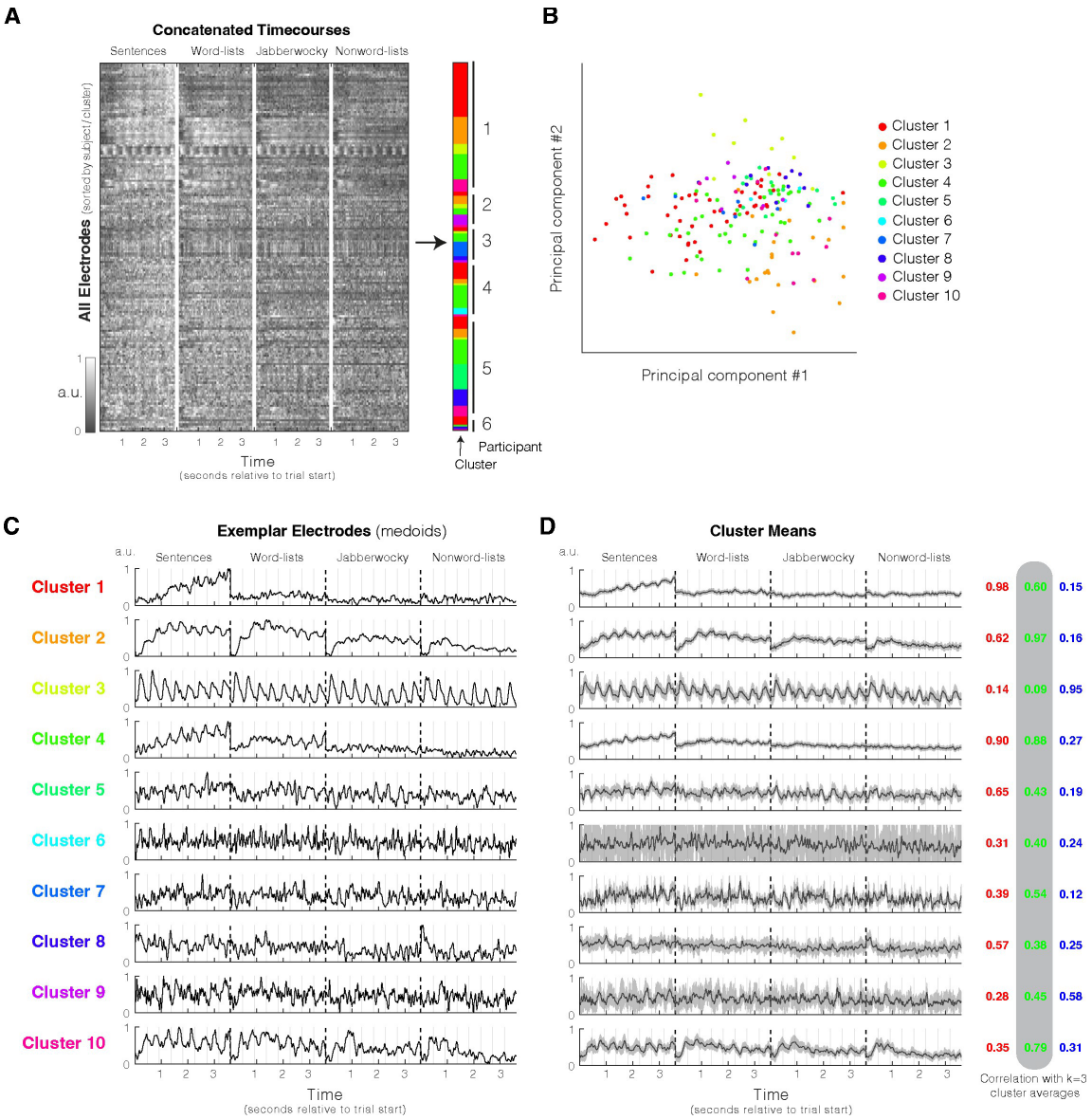


Figure R15 (Figure S2 in Supplementary Information). Dataset 1 k-medoids clustering with k=10. **A)** Clustering mean electrode responses (S+W+J+N) using k-medoids (k=10) with a correlation-based distance. Shading of the data matrix reflects normalized high-gamma power (70-150Hz). **B)** Electrode responses 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. Many of the additional clusters resemble the main three response profiles from **Figure 2**.

* The use of a [1, -1] ideal label assignment for the sentence vs. nonword lists, in order to identify language responsive electrodes, is maybe not the best, given that the high gamma signal can never be below zero. It might be good to check that an assignment of [1, 0] gives the same result.

Thank you – we have tried your suggested version of the assignment in our earlier version of these analyses, and it produces the same results.

* Minor thing but I recommend avoiding saying “an electrode processes information” and instead more precisely talk about neural populations as captured by the electrode

Fixed, thank you. (We still use this shorthand in places, to minimize verbosity, but we have explicitly stated what we mean when we use this description in the beginning of the Discussion section.)

Reviewer #3:

[Summary omitted]

Overall, the study’s main finding is interesting and enhances our understanding of the distribution of neural populations with different temporal integration periods. The authors also show good robustness for this finding across datasets. However, the current analyses focus too much on clusters rather than on individual electrode properties inside each cluster. I have few major concerns:

1. The TRW modeling analysis is intuitive, but it doesn’t add much in its current state because the average time courses of each cluster seem to show the same information already because the analysis only discusses the results at the cluster level, rather than the **individual electrode level.** Cluster 1 and cluster 2 have very wide distributions of fitted TRW (fig 4C), and it would have been nice to analyze this more and look at the electrodes in these groups that were on the far ends of the distribution to look at what about them put them in their respective cluster despite the TRW model choosing a very different receptive field. **This whole analysis could be significantly enhanced if, rather than just looking at each cluster, each electrode’s best fitted receptive window was used to color a visualization of the electrodes on the brain, which may help show some interesting gradients of TRW along different processing pathways**. While the authors seem to claim that these clusters of population responses illustrate discrete types of responses throughout the language network, that doesn’t mean that the TRW of populations doesn’t follow a more continuous gradient.

Thank you for this comment and a helpful suggestion. We have now generated the anatomical **Figure R16** (panel **F** in **Figure 5** of the revised manuscript) below, in which each electrode is colored by its estimated TRW. We think that this figure is indeed helpful, as it shows the general anatomical trend of estimated TRWs, not tied to the discrete cluster assignment, and makes apparent a general trend of TRW size increasing in the posterior to anterior direction. However, there are numerous exceptions - some electrodes with short TRWs fall in anterior frontal regions, and some electrodes with long TRWs fall in posterior temporal regions. This visualization thus aligns with our general finding of a gross anatomical trend, together with a local mosaic pattern. We also acknowledge (and have made this acknowledgement clearer in the revised manuscript) that we cannot rule out the possibility of functional gradients at a finer spatial scale, as we are limited by the sparsity of the ECoG/sEEG coverage. However, our current results

indicate that at the macro-anatomical level there are no obvious gradients beyond the posterior-to-anterior trend mentioned above.

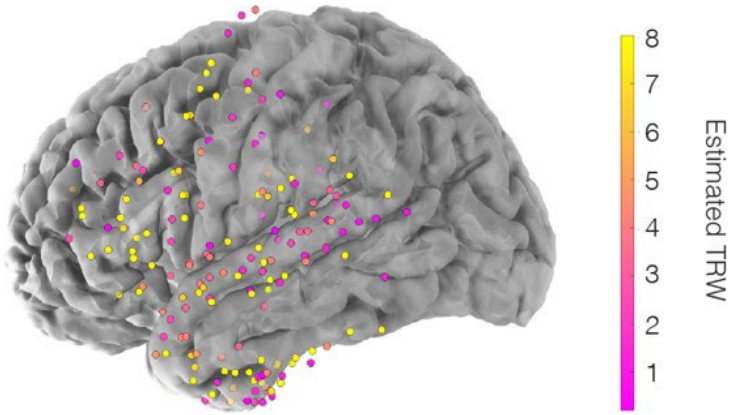
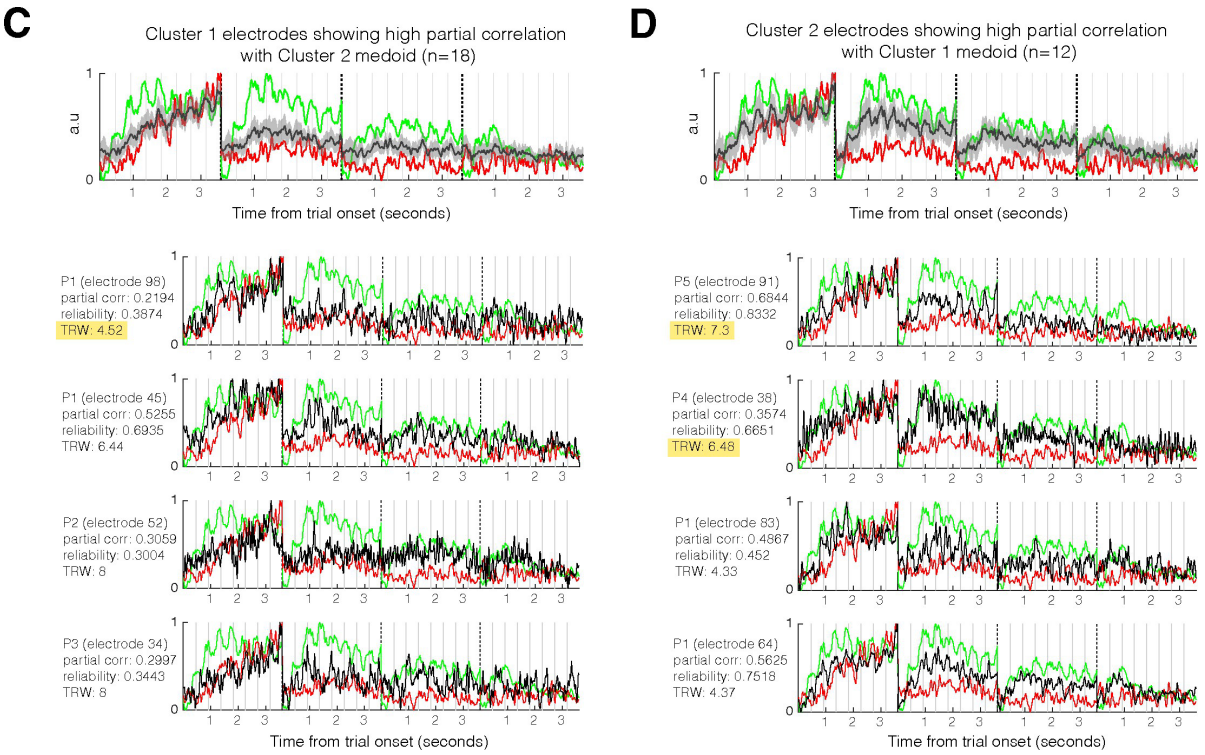


Figure R16 (panel F of Figure 5 in the revised manuscript). Anatomical distribution of electrodes in Dataset 1 colored by their estimated temporal receptive window (TRW, **Figure 4** in the revised manuscript). There is a slight trend of increasing TRW size from posterior to anterior regions, but considerable local heterogeneity exists.

In an effort to further address your comment, we have also attempted to explore individual response profiles (see our response to Reviewer #2 for a related discussion). As you point out, there are some electrodes that belong to Cluster 1 and have a relatively short TRW or vice versa – electrodes in Cluster 2 that have a long TRW. When we examine the response profiles of such electrodes (see the highlighted electrodes in **Figure R14C-D** from our reply to Reviewer #2, duplicated and highlighted in yellow below for convenience), we find a couple features that contribute to these electrodes being assigned a TRW that differs from what would be expected by their cluster assignment. Namely, for Cluster 1 electrodes assigned a short TRW (e.g., P1 electrode 98, **Figure R14C**), their responses to the Sentence condition exhibit a faster buildup of activity than other Cluster 1 electrodes. In contrast, for Cluster 2 electrodes assigned a long TRW (e.g., P5 electrode 98, P4 electrode 38, **Figure R14D**), their responses to the Sentence condition exhibit a more gradual buildup of activity than other Cluster 2 electrodes, and no plateauing (or even decaying) of activity, which is present in some Cluster 2 electrodes (however, note that their responses to the other conditions *do* look more like typical Cluster 2 responses). As we discuss above, these response profiles display features that are characteristic of both Cluster 1 and Cluster 2.
This indicates either that a continuum of responses between the cluster medoids is, to some degree, present in the observed neuronal responses, or that some electrodes are capturing responses from multiple neural populations. In order to adjudicate between these two alternatives, a higher spatial resolution is needed. However, although we find some evidence for mixed responses, we also note that most of the high reliability electrodes in each cluster are assigned approximately the same TRW. We have now added this issue to the Discussion.



Panels C and D from Figure R14, duplicated here for convenience (Figure S4 in the Supplementary Information). *Responses from electrodes with a TRW size that differs from their assigned cluster's TRW size are highlighted in yellow.* **C)** Response profiles from electrodes assigned to Cluster 1 that have a high partial correlation (>0.2, arbitrarily defined) with the Cluster 2 medoid (and split-half reliability>0.3). *Top:* Average over all electrodes that meet 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 have a high partial correlation (>0.2, arbitrarily defined) 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).

2. The following argument needs to become more compelling; since the more-reliable (i.e. less noisy) electrodes show clearer separation among clusters in low-dimensional space, then this is evidence that these clusters characterize the structure of responses and argue against a uniform continuum of response types. If the 2D Principal components plot (fig 2C) shows the less reliable electrodes towards the center, couldn't that be because the less reliable electrodes have an average time course which gets washed out more after averaging, so the principal components go towards the average more, rather than towards the clusters? If so, this figure doesn't seem like it's evidence that there isn't a continuum, because this should naturally happen to less reliable electrodes. If these principal components come from average time-courses, then they are naturally tied to the reliability metric of correlation between even and odd trials.

We understand your concern about using the PCA space to support cluster separation and not a uniform continuum of response types, and we have now removed this argument from the revised manuscript. Instead, in an effort to strengthen the claim that our data is well described by k=3 clusters, we now performed an additional analysis where we clustered Dataset 1 with k ranging from 1 to 10 while parametrically excluding some percentage of the electrodes due to their lower reliability. When we do this (see **Figure R5** from above, duplicated below for convenience, and inset in **Figure 2A** in the manuscript), we find that the elbow at k=3 becomes stronger as more unreliable electrodes are excluded. That is, the amount of additional explained variance when clustering with values of k>3 is smaller when clustering only high-reliability electrodes. We believe that this evidence strengthens our claim that there exist ***three*** dominant response profiles within the language-sensitive neural populations in our data. However, we cannot rule out the possibility of an underlying continuum of responses (see **Figure R14** above for additional discussion), and we better acknowledge this limitation in the Discussion.

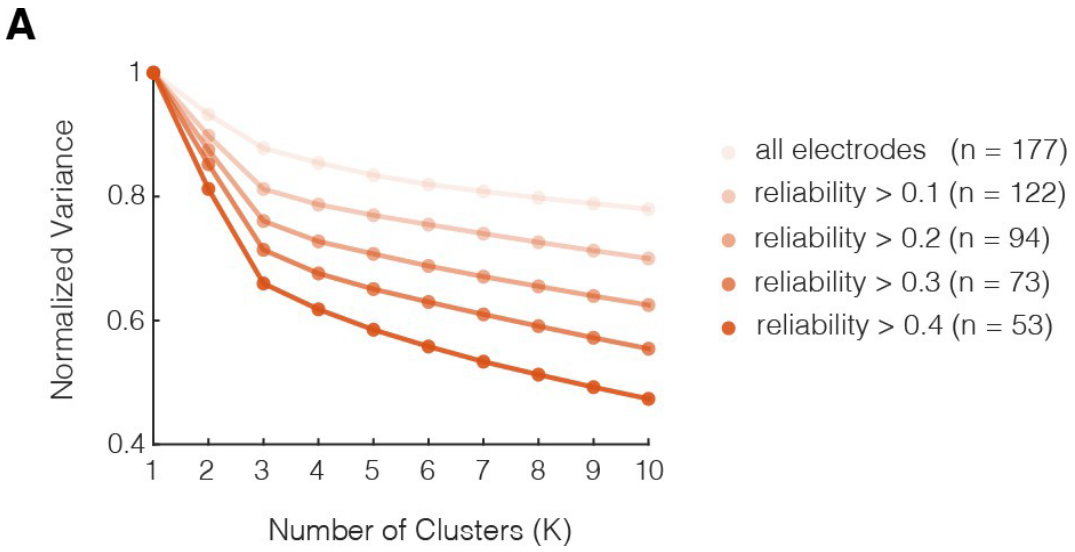


Figure R5, duplicated here for convenience (inset for Figure 2A in the revised manuscript). Clustering while omitting electrodes below a parametrically sampled reliability threshold. Normalized variance is presented against the number of clusters (K), to illustrate the elbow method. The elbow gets more pronounced when eliminating lower-reliability electrodes, which strengthens our claim that k=3 best describes these data.

3. Another critical factor is the issue of subject variability. While the same clusters can appear in different subjects, this is not enough to claim that all subjects have the same response property. For example, one can see from Figure 3B that cluster 1 only shows a clear incremental response to sentences in the top subject. While this figure indicates that cluster 1 also appears in different subjects, it is not clear from this figure if other subjects (red rows) show the same incremental pattern (at least not from 3B). Again, the problem with clustering is that this analysis forces a categorical label upon electrodes, based on a comparison of a similarity metric. However, if an electrode belongs to a cluster, it doesn't necessarily mean that electrode does the same thing as the mean of the cluster. What would be helpful is to include the **average**

cluster plots (3E) for individual subjects and show they all have the same response pattern across the four stimuli.

Thank you for the chance to provide more information about the individual participants' data. **Figure R18** below (and **Figure S1** in the revised manuscript) shows the average cluster responses for the six participants in Dataset 1. Although the buildup is most salient in P1 (who contributes the most electrodes overall as well), the buildup is present in all other participants (with varying degrees of steepness), in line with several groups having now replicated the build-up effect during sentence comprehension (Nelson et al., 2017 *PNAS*; Desbordes et al., 2023 *J Neurosci*; Woolnough et al., 2023 *PNAS*), which was first reported by our group in 2016 using the participants from Dataset 1 (Fedorenko et al, 2016 *PNAS*). Cluster 2 also shows good correspondence across participants (except for P6 who only had a single electrode assigned to this cluster). Cluster 3 is the least visually consistent across participants, plausibly because it has the fewest electrodes.

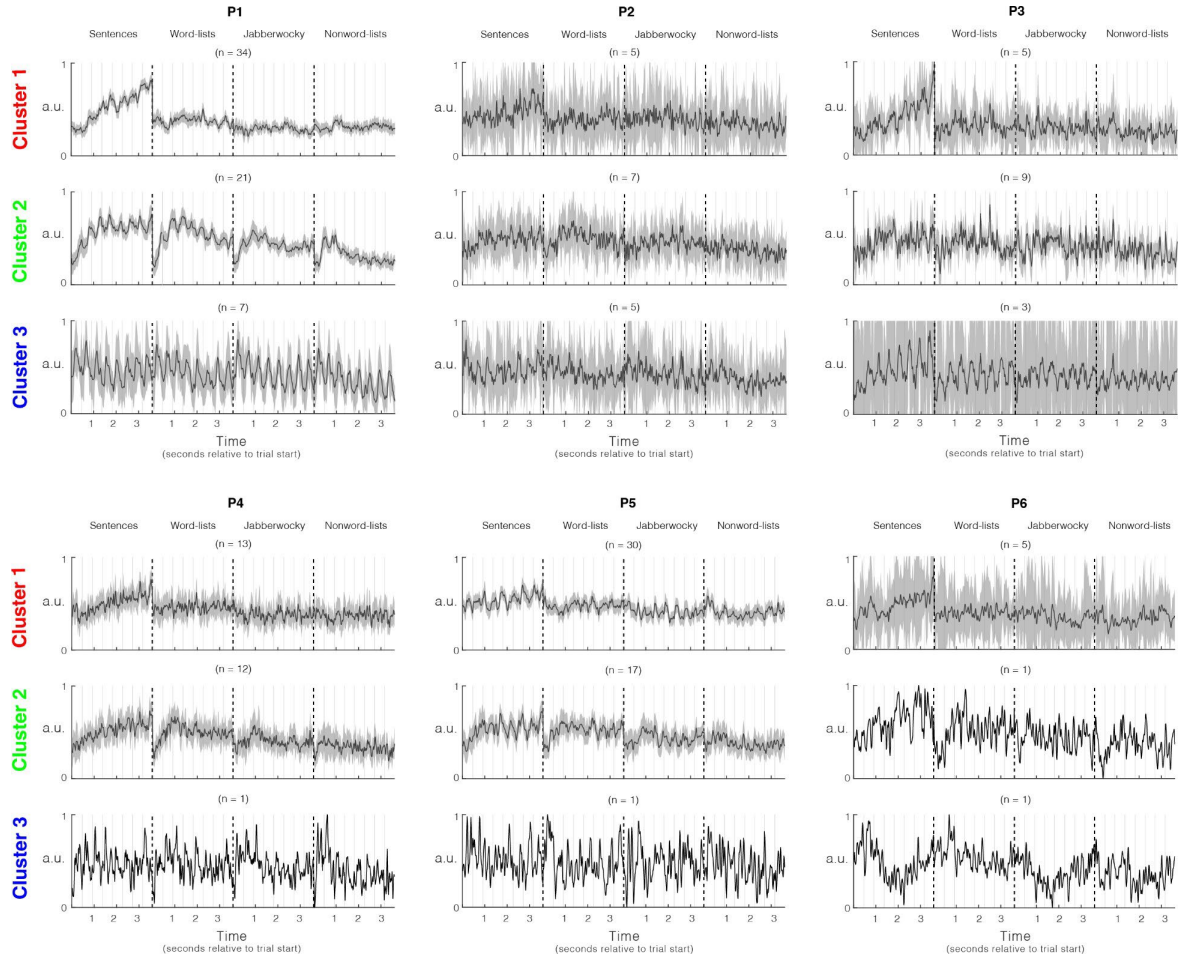


Figure R18 (Figure S1 in Supplementary Information). Dataset 1 k-medoids (k=3) cluster assignments by participant. Average cluster responses as in **Figure 2E** of the revised manuscript grouped by participant. Shaded areas around the signal reflect a 99% confidence interval over electrodes. The number of electrodes constructing the average (n) is denoted above each signal in parenthesis. Prototypical responses for each of the three clusters

are found in nearly all participants individually. However, for participants with only a few electrodes assigned to a given cluster (e.g., P5 Cluster 3), the responses can be more variable.

Minor points:

1. Please add more details about the stimuli. I understand that stimulus information is available in previous studies, but this paper needs to be self-contained.

Thank you – added in the Methods.

2. The citations are pretty narrow and rely heavily on self-referencing. This needs to be expanded. Several relevant recent papers have looked at similar questions and showed a similar distributed pattern of linguistic processing, including Ding et al. 2016, de Heer et al. 2017, Keshishian et al. 2023, and Brodbeck et al., 2018

Thank you – Keshishian et al. 2023 as well as several other references have been added throughout the manuscript (see the blue entries in our reference list to identify all new references). We omitted references which we did not find directly relevant to our study.

3. I do not think the first figure is important enough to be included as main and can be moved to supplemental, discussion, or removed entirely. The paper only uses ECoG, which has been used for decades to study speech in humans. There is no need to have a main figure that discusses its superior spatial resolution compared to other recording techniques.

We agree – removed.

Decision Letter, first revision:

24th February 2024

Dear Dr Regev,

Thank you once again for your revised manuscript, entitled "Neural populations in the language network differ in the size of their temporal receptive windows," and for your patience during the re-review process.

Your manuscript has now been evaluated again by Reviewer #1 and #3. Reviewer #2 was unable to re-review, so we asked the other two reviewers to comment on your responses to Reviewer #2's concerns; a summary of their responses is included below alongside the other reviewer comments.

Although the reviewers found your manuscript to have improved during revision, they also raise some important outstanding concerns (Reviewer #1 and one issue outstanding in regards to Reviewer #2). We remain interested in the possibility of publishing your study in Nature Human Behaviour, but would like to consider your response to these outstanding concerns in the form of a revised manuscript before we make a decision on publication.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

In sum, we invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We hope to receive your revised manuscript within 4-8 weeks. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

• Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors and reviewers to evaluate your revision.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Sincerely,

[REDACTED]

REVIEWER COMMENTS:

Reviewer #1:
Remarks to the Author:
Dear Tamar, Colton, and Co-authors,

Congratulations on producing this thorough, responsive, and transparent revision to the initial submission of your manuscript. I particularly appreciate your: (i) inclusion of figure S3, which provides a complete and transparent description of Data Set 1; (ii) exploration of the effects of electrode reliability on the clustering solution and the LMEs; (iii) more rigorous implementations of statistical inference for the linear decoder and LMEs; (iv) thorough expansion of the toy model to rule out effects of kernel shape; and (v) inclusion of anatomical maps showing partial cluster correlations (Fig. S4E) and estimated TRWs (Fig. 5F). I was also encouraged to see that the clustering solution held up after z-scoring electrode time series within condition. Most of my original concerns have been addressed. I do have some lingering minor concerns that I will break into two categories: those that, in my opinion, deserve some attention, and those stated only for completeness.

Minor concerns with revisions suggested:

1. In response to my original concern about functional heterogeneity in Cluster 2 and, in particular, the (partial) failure of Cluster 2 to replace in Data Set 2, the authors performed an additional analysis in which electrodes from Data Set 2 were assigned "cluster" labels based on maximum correlation with the cluster medoids estimated from Data Set 1. Two statistical inferences were then implemented using these cluster labels: (i) a permutation test comparing cluster similarity across Data Sets 1 and 2; and (ii) an LME testing

the effect of cluster on TRW within Data Set 2. Each of these seem to be a case of "double dipping" – that is, the "clusters" in Data Set 2 were pre-selected to be similar to those in Data Set 1. I think we can therefore reasonably assume both statistical inferences listed above are subject to bias (though there may be something I am missing about the permutation test that avoids such bias). In my opinion, it would be appropriate to assign cluster labels using this approach, which is essentially a "winner take all" (WTA) relative to the correlation of each electrode with the Data Set 1 medoids, and then report the number of electrodes from Data Set 2 falling into each WTA category (as is currently reported for Cluster 2). However, subsequent statistical inferences based on the WTA labels should be avoided (e.g., those related to Figs. S8 and S10). Even if the WTA is omitted entirely from the manuscript, the authors have been fully transparent elsewhere about the potential overlap between Clusters 1 and 2 (e.g., Fig. S4), which is a satisfactory response to my original concern. Readers will have enough information to decide for themselves whether the clustering solution is plausible/believable.

2. A very minor comment on the LMEs comparing TRWs in Data Set 1 across subsets of participants with fast vs. slow stimulus presentation rates. As the authors note, when separate LMEs are run on each subset (n = 3 per subset), there are few denominator DFs left to estimate p-values with Satterthwaite correction. An alternative may be to fit a single LME with a between-subject effect of presentation rate and a within-subject effect of cluster. For the subset analysis, the effect of interest is the interaction between the two. If the main effect of presentation rate is omitted from the model, the interaction is then the simple effect of cluster within presentation rate.

3. Assorted minutiae. (A) In the abstract, consider retaining the description of the TRW model as a "toy model" to be consistent with the main manuscript the response to Reviewer 2. (B) Revised manuscript p. 21, first paragraph: there are two consecutive sentences beginning with "Although." (C) Revised manuscript pp. 22-23, first paragraph of Future Directions: you have a "First, Second" construction nested within another "First, Second" construction, disrupting flow of the paragraph. (D) Revised manuscript p. 34: I'm not sure if this was an artifact of building the PDF, but somehow the page number showed up in the middle of a word: "This approach is crucial f34or accommodating inter-individual variability."

Comments stated only for completeness:

1. I was not convinced by some of the responses to Reviewer 2 related to item-level variability. In particular, Fig. R7 in the rebuttal. There are two sources of bias in this analysis. First, electrodes were pre-sorted into groups with similar response profiles (i.e., clusters), so we might expect positive item-to-condition-average correlations due to some intrinsic response property of the neuronal populations underlying the electrodes in a given cluster. Second, each item itself contributes to the condition average (information leak). Thus, we might expect some degree of positive correlation due only to bias. More important may be the spread of these correlations around their means, whose magnitude suggests the possibility of nontrivial item-level variance. I also wonder, in Fig. R8, why the spread of within-item correlations is larger than that of between-item correlations (even though each is centered on approximately zero). Without more detail on the R8 analysis, this is difficult to determine. However, I agree with the authors' other arguments related to item-level variance: (i) the study was not designed to measure it; (ii) the stimuli are not amenable to such measurement by chance; and (iii) the data are probably too noisy to allow robust item-level measurements.

2. Reviewer 3 raised an important point about subject-level variance. The context was related to whether the clustering solution would hold across subjects. I would extend this to note that the anatomical distribution of electrodes belonging to each cluster is strongly driven by those subjects with more electrodes/better coverage. Indeed, when considered alone, P1 appears to demonstrate a robust spatial correlation among electrodes by cluster label (Fig. 5B). It is possible the spatial "mosaic" of clusters observed across all electrodes is an artifact of combining data across subjects. This is partially addressed by the LMEs estimating the effect of cluster on the Y and Z coordinates of the electrodes, which capture participant-level variability by including a random effect of cluster. However, the random cluster effect is, by constraint of electrode placement per subject, not an "apples to apples" comparison across subjects, making the subject-level variance of the effect difficult to estimate/interpret. Nevertheless, when the LMEs are weighted by electrode reliability, there seems to be a reliable spatial division of the clusters (Fig. 5C). Taking these observations together with some additional hints of macro-scale spatial organization in Figs. 5F and S4E, I'm still not completely on board with the interpretation that the clusters are interleaved in

local patches throughout the language network. However, readers can again come to their own conclusion.

Overall, this work is very strong and I commend the authors for their attention to detail and transparency.

All the best,

Jon Venezia

Reviewer #2:
SUMMARY OF COMMENTS MADE BY OTHER REVIEWERS

Reviewer #2 raised concerns regarding the need for more investigation into stimulus variability. The point here is to rule out the possibility that the patterns observed might have been due to a small set of trials in each condition (for an unknown reason) and not directly related to the condition. The authors respond by providing evidence that the main findings are robust to inter-item variability, basically taking a random sample response and comparing it to the average time course. Doing this 100 times results in positive correlations, which is interpreted as evidence for robustness to stimulus variabilities.

One issue with this interpretation is that it is unclear whether this result can rule out the null hypothesis, given the lack of statistical tests (e.g., is an r-value of 0.3 good enough compared to a random assignment of conditions that could also have a high correlation due to some unknown effect?).

A complementary and more straightforward analysis showing trial-level responses, similar to how the authors present all electrodes, can be helpful. Such a figure would look like Fig. 2B, but the y-axis could be trials sorted by conditions, where each row shows the average response of each cluster to that trial. While qualitative, this figure will hopefully show that most trials in the same condition evoke a similar average response. The authors further argue why this stimulus is not optimal for stimulus-level analysis of higher-order selectivity, but I don't think that's what Reviewer #2 was asking. Instead, the critique questions the robustness of effects on stimulus variability.

Reviewer #3:
Remarks to the Author:
The authors have done a great job addressing the concerns. I have no further comments.

Author Rebuttal, first revision:

REVIEWER COMMENTS:

Reviewer #1:

[summary omitted]

1. In response to my original concern about functional heterogeneity in Cluster 2 and, in particular, the (partial) failure of Cluster 2 to replace in Data Set 2, the authors performed an additional analysis in which electrodes from Data Set 2 were assigned "cluster" labels based on maximum correlation with the cluster medoids estimated from Data Set 1. Two statistical inferences were then implemented using these cluster labels: (i) a permutation test comparing cluster similarity across Data Sets 1 and 2; and (ii) an LME testing the effect of cluster on TRW within Data Set 2. Each of these seem to be a case of "double dipping" – that is, the "clusters" in Data Set 2 were pre-selected to be similar to those in Data Set 1. I think we can therefore reasonably assume both statistical inferences listed above are subject to bias (though there may be something I am missing

about the permutation test that avoids such bias). In my opinion, it would be appropriate to assign cluster labels using this approach, which is essentially a “winner take all” (WTA) relative to the correlation of each electrode with the Data Set 1 medoids, and then report the number of electrodes from Data Set 2 falling into each WTA category (as is currently reported for Cluster 2). However, subsequent statistical inferences based on the WTA labels should be avoided (e.g., those related to Figs. S8 and S10). Even if the WTA is omitted entirely from the manuscript, the authors have been fully transparent elsewhere about the potential overlap between Clusters 1 and 2 (e.g., Fig. S4), which is a satisfactory response to my original concern. Readers will have enough information to decide for themselves whether the clustering solution is plausible/believable.

Thank you for elaborating on this point. We do recognize the potential for bias in our statistical approach (“winner take all”, WTA) given that the criterion that we evaluated was the same as (when testing cluster similarity)—or related to (when testing TRW)—the criterion that we optimized for. Note, however, that we had performed the **exact same optimization** on the surrogate data permutations to which we compared the real data. That is, we assigned permuted electrodes to a group due to their similarity to the clusters from Dataset 1 in the same way as we did with the real data, which ensures that the bias is also present in the null distribution. We do agree though that statistics are not essential to make this point (the apparent visual similarity of the WTA groups from Dataset 2 to the clusters from Dataset 1 demonstrates that Cluster-2-like responses *do* exist in Dataset 2). We have therefore removed the first of these tests from the paper, as you suggest. However, we believe that the LME model that tests the effect of cluster on TRW is still valid given that the variable that we evaluate (TRWs) was not optimized for in the WTA procedure, and we would prefer to keep it in the manuscript.

Additionally, we adopted your suggestion to report the number of electrodes assigned to each group in the WTA approach for Dataset 2 (**Figure RR1** below, also added to **Figure S8** in the manuscript), which revealed that a non-trivial number of electrodes belonged to Group 2 (the analog of Cluster 2, n=95, **Figure RR1A**). We also report the *difference* in the number of electrodes assigned to each cluster with the WTA vs. the clustering procedure (**Figure RR1A**). Visualizing the average timecourse of electrodes grouped by their assignment using the WTA vs. the clustering approach (**Figure RR1B**), revealed that the electrodes that exhibited the *most* Cluster-2-like response were initially assigned to Cluster 1 under the clustering procedure but were “pulled out” into Group 2 using the WTA approach (n=44; **Figure RR1B**). Because this figure seems useful for readers, we have included it in **Figure S8** of the Supplementary Information.

We also think the description of the analysis as a “winner-take-all” approach is particularly easy for readers to understand and now refer to it that way in the manuscript, so thank you for that suggestion.

2. A very minor comment on the LMEs comparing TRWs in Data Set 1 across subsets of participants with fast vs. slow stimulus presentation rates. As the authors note, when separate LMEs are run on each subset (n = 3 per subset), there are few denominator DFs left to estimate p-values with Satterthwaite correction. An alternative may be to fit a single LME with a between-subject effect of presentation rate and a within-subject effect of cluster. For the subset analysis, the effect of interest is the interaction between the two. If the main effect of presentation rate is omitted from the model, the interaction is then the simple effect of cluster within presentation rate.

Thank you for this suggestion. We ran the model you suggested, and the results (**Tables RR1** and **RR2** below) show that there was no significant overall interaction between cluster and presentation rate. More specifically, the presentation rate affected only responses in Cluster 1 but not in Clusters 2 and 3 (**Table RR1**), and the overall interaction as measured by ANOVA (**Table RR2**) did not reach significance. We have updated **Tables S6-7** in the manuscript accordingly.

3. Assorted minutiae. (A) In the abstract, consider retaining the description of the TRW model as a “toy model” to be consistent with the main manuscript with the response to Reviewer 2. (B) Revised manuscript p. 21, first paragraph: there are two consecutive sentences beginning with “Although.” (C) Revised manuscript pp. 22-23, first paragraph of Future Directions: you have a “First, Second” construction nested within another “First, Second” construction, disrupting flow of the paragraph. (D) Revised manuscript p. 34: I’m not sure if this was an artifact of building the PDF, but somehow the page number showed up in the middle of a word: “This approach is crucial f34or accommodating inter-individual variability.”

Thank you for these comments. We have corrected these concerns in the manuscript. The only exception was for point (A) where we feel like the term “toy model” is not clear enough to appear in the abstract, whereas only in the manuscript we have more space to further explain what we mean by “toy model”.

Comments stated only for completeness:

1. I was not convinced by some of the responses to Reviewer 2 related to item-level variability. In particular, Fig. R7 in the rebuttal. There are two sources of bias in this analysis. First, electrodes were pre-sorted into groups with similar response profiles (i.e., clusters), so we might expect positive item-to-condition-average correlations due to some intrinsic response property of the neuronal populations underlying the electrodes in a given cluster. Second, each item itself contributes to the condition average (information leak). Thus, we might expect some degree of positive correlation due only to bias. More important may be the spread of these correlations around their means, whose magnitude suggests the possibility of nontrivial item-level variance. I also wonder, in Fig. R8, why the spread of within-item correlations is larger than that of between-item correlations (even though each is centered on approximately zero). Without more detail on the R8 analysis, this is difficult to determine. However, I agree with the authors’ other arguments related to item-level variance: (i) the study was not designed to measure it; (ii) the stimuli are not amenable to such measurement by chance; and (iii) the data are probably too noisy to allow robust item-level measurements.

Thank you for your thorough evaluation of the comments of the other reviewers. We understand the concerns you raise regarding the item-level variability. The first point concerns the analysis we presented in **Figure R7** of our previous revision, which evaluated the robustness of our results to inter-item variability. As you point out (and as Reviewer #3 also commented below), the correlation values presented in the **Figure R7** histograms are hard to interpret and we therefore removed this figure from OSF (https://osf.io/xfbr8/). Instead, we performed two additional analyses to further establish that our results do not depend on a small number of trials. First, we provide a new visualization of item-level variability, as suggested by Reviewer #3 below (**Figure RR2** and **Figure RR3**; see our responses to Reviewer #3 below for details), which transparently presents the responses to all conditions and in all clusters by item. Additionally, we present a new analysis where we recompute

the average response in each cluster after removing progressively more individual items (**Figure RR4**; also see our responses to Reviewer #3 below). Both analyses further demonstrate the similarity of responses across items, and we have included them on OSF.

Your second point concerns the larger spread of the within-item correlations than the between- item correlations (**Figure R8** of the previous revision). This difference resulted from the way that the averaging across individual item-to-item correlations was performed. In particular, there are more possible between-item comparisons (n*(n-1)/2 comparisons for n items) than within-item comparisons (n comparisons for n items). In the previous version, we had constructed all the histograms from exactly n points (for both the between-item and within-item histograms). In the between-items case, to compute a single value per item, we had averaged the correlation values of a given item with all the other items. This averaging shrank the variance of the between-item distribution. We have now fixed this issue by computing the histograms using the correlation values of *all* the relevant item pairs without averaging (**Figure RR5** below, which we have added as **Figure O8** on OSF. For transparency, we have also added to OSF a visualization of these correlation values *by item*, **Figures O9-10**). As a reminder, the point of this figure was to demonstrate that responses within items were not more similar than responses between items, suggesting that the stability of responses per item was not high enough relative to the measurement noise to make analyses investigating effects of item properties on neural responses meaningful.

2. Reviewer 3 raised an important point about subject-level variance. The context was related to whether the clustering solution would hold across subjects. I would extend this to note that the anatomical distribution of electrodes belonging to each cluster is strongly driven by those subjects with more electrodes/better coverage. Indeed, when considered alone, P1 appears to demonstrate a robust spatial correlation among electrodes by cluster label (Fig. 5B). It is possible the spatial "mosaic" of clusters observed across all electrodes is an artifact of combining data across subjects. This is partially addressed by the LMEs estimating the effect of cluster on the Y and Z coordinates of the electrodes, which capture participant-level variability by including a random effect of cluster. However, the random cluster effect is, by constraint of electrode placement per subject, not an "apples to apples" comparison across subjects, making the subject-level variance of the effect difficult to estimate/interpret. Nevertheless, when the LMEs are weighted by electrode reliability, there seems to be a reliable spatial division of the clusters (Fig. 5C). Taking these observations together with some additional hints of macro-scale spatial organization in Figs. 5F and S4E, I'm still not completely on board with the interpretation that the clusters are interleaved in local patches throughout the language network. However, readers can again come to their own conclusion.

We understand your concern and we have revised the text in the manuscript to acknowledge alternative interpretations (see Discussion). It is true that there is substantial participant-level variance in the anatomical distribution of electrodes, however, given the small number of participants and the substantial degree of variability in electrode coverage, we are hesitant to make strong claims about the macro-scale anatomical organization of these neural populations that was only present in a subset of participants. As a result, we softened any anatomical claims made in the text, and explicitly acknowledged the possibility that a macro-scale organization could become more evident with more participants and a more systematic coverage of the frontal and temporal cortex (Discussion).

Reviewer #2: SUMMARY OF COMMENTS MADE BY OTHER REVIEWERS

Reviewer #2 raised concerns regarding the need for more investigation into stimulus variability. The point here is to rule out the possibility that the patterns observed might have been due to a small set of trials in each condition (for an unknown reason) and not directly related to the condition. The authors respond by providing evidence that the main findings are robust to inter- item variability, basically taking a random sample response and comparing it to the average time course. Doing this 100 times results in positive correlations, which is interpreted as evidence for robustness to stimulus variabilities.

One issue with this interpretation is that it is unclear whether this result can rule out the null hypothesis, given the lack of statistical tests (e.g., is an r-value of 0.3 good enough compared to a random assignment of conditions that could also have a high correlation due to some unknown effect?).

A complementary and more straightforward analysis showing trial-level responses, similar to how the authors present all electrodes, can be helpful. Such a figure would look like Fig. 2B, but the y-axis could be trials sorted by conditions, where each row shows the average response of each cluster to that trial. While qualitative, this figure will hopefully show that most trials in the same condition evoke a similar average response. The authors further argue why this stimulus is not optimal for stimulus-level analysis of higher-order selectivity, but I don't think that's what Reviewer #2 was asking. Instead, the critique questions the robustness of effects on stimulus variability.

Thank you very much for your willingness to comment on our reply to another reviewer. And thank you for the great suggestion. Below are the figures you proposed, which qualitatively show that most trials in the same condition evoke a similar response (**Figure RR2** and **Figure RR3**). We hope you find these figures compelling as a demonstration that our results are not driven by a small number of items and are robust to inter-item-variability.

To further test the robustness of our findings to inter-trial variability, we performed an additional analysis to evaluate the sensitivity of the clusters to item loss. Namely, we re-computed the average cluster responses after randomly omitting progressively more items (**Figure RR4**). Specifically, we a) randomly removed *n* items from the dataset, where *n* ranged from 0 to 52, the total number of unique items that were seen by all participants, b) re-computed the average response of each *electrode*, c) re-computed the average response of each *cluster* (using the original electrode assignments), and finally d) compared the resulting responses with the responses of clusters based on all items (**Figure RR4**). We repeated the random sampling of items 100 times for every value of *n*. This analysis further suggests that individual items have a relatively minimal impact on the average cluster response. Large drops in similarity are not observed until 30+ items are removed. Thus, our condition-level findings are not driven by a small number of items. We think **Figures RR2-4** may be helpful for readers and have therefore included them on OSF (https://osf.io/xfbr8/). Please also check our related reply to Reviewer #1 above ('Comments stated only for completeness, 1').

Reviewer #3:

The authors have done a great job addressing the concerns. I have no further comments. Thank you for the positive evaluation and helping us strengthen the manuscript.

Figures

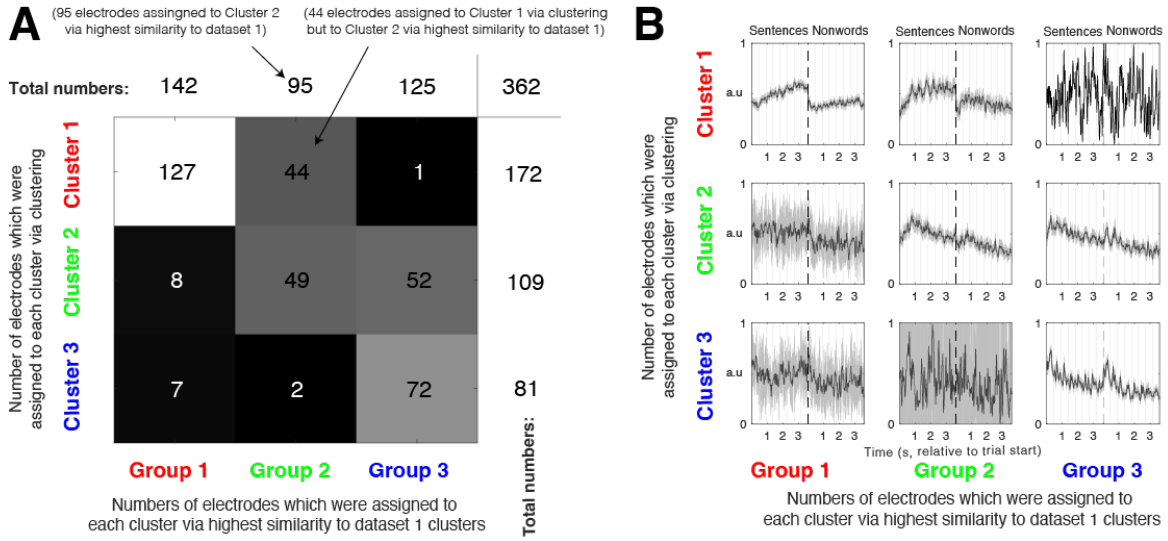
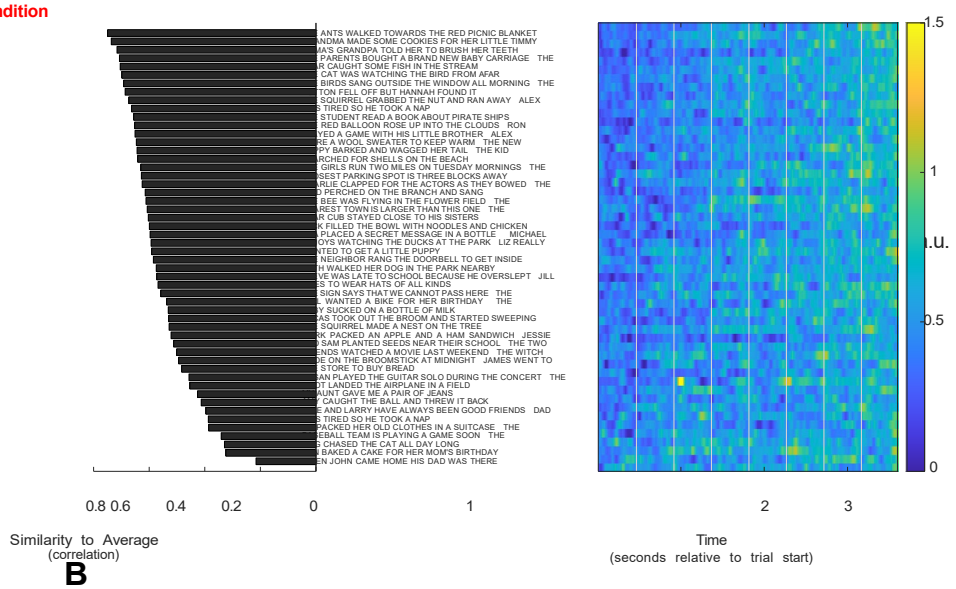


Figure RR1 (added to Supplementary **Figure S8** in the manuscript). Comparison of cluster assignment of electrodes from Dataset 2 using clustering vs. winner-take-all (WTA) approach. **A)** The numbers in the matrix correspond to the number of electrodes assigned to cluster y during *clustering* (y-axis) 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 numbers of electrodes assigned to each cluster during the clustering approach are shown to the right of each row. The total numbers 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. **B)** Similar to A, but here the average timecourse across all electrodes assigned to the same cluster/group during both procedures is presented. Shaded areas around the signals reflect a 99% confidence interval over electrodes.

A

Cluster 1: S condition

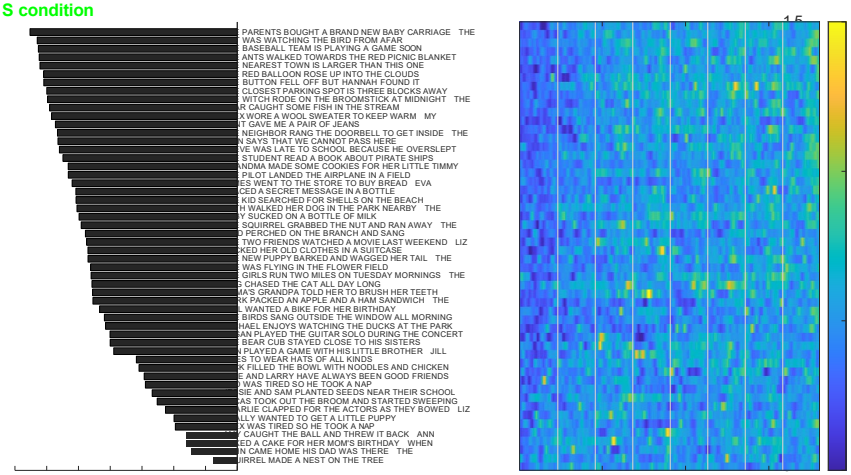


Similarity to Average (correlation)

Time (seconds relative to trial start)

B

Cluster 2: S condition

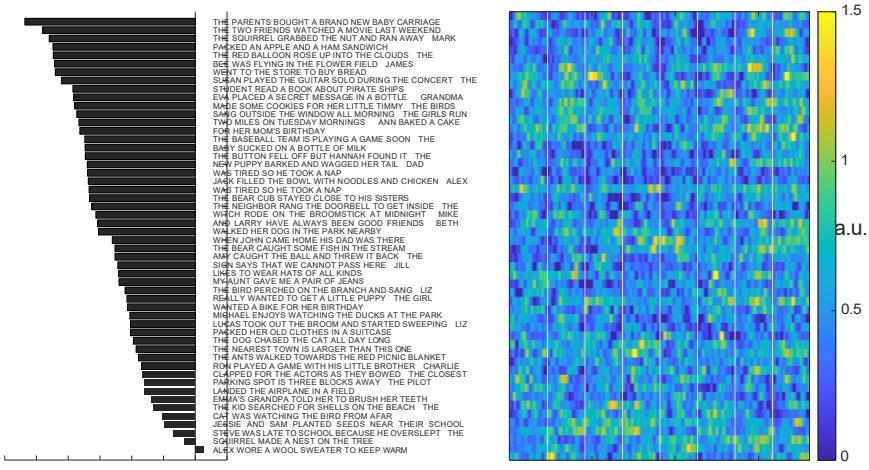


Similarity to Average (correlation)

Time (seconds relative to trial start)

C

Cluster 3: S condition

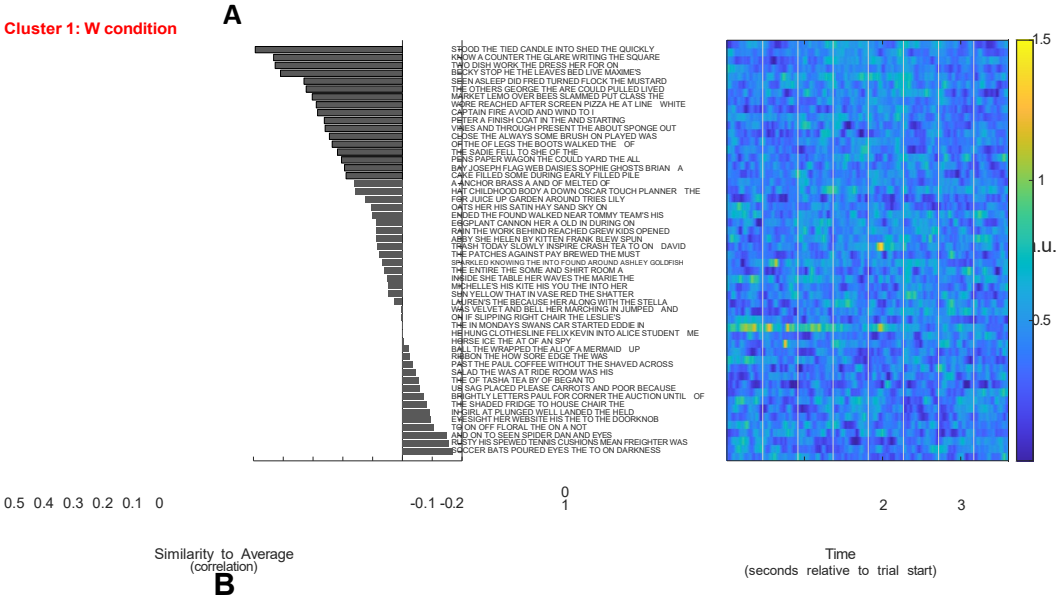


Similarity to Average (correlation)

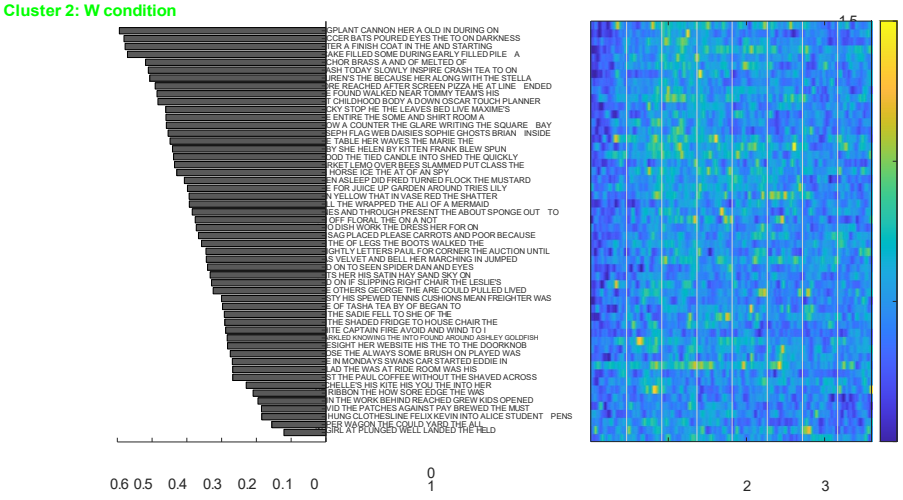
Time (seconds relative to trial start)

Figure RR2 (added as OSF **Figure O5**). Responses by cluster to individual items from the Sentence condition. For each cluster, items are sorted by their similarity to the condition average (excluding the item being evaluated to avoid information leak), and the normalized high gamma power (a.u.) of each item, averaged over all electrodes in a given cluster (Cluster 1, **A**; Cluster 2, **B**; Cluster 3, **C**), is shown in the heatmap. Only the items that were shown to all participants (n=52) are displayed. Electrode responses to individual items were normalized by the same factors that were used to normalize the electrode's average response over all items (such that the average over all item timecourses for a given electrode will yield the same average electrode response profile that was used in other analyses). This figure demonstrates that individual items within the Sentence condition evoked similar responses.

Cluster 1: W condition



Cluster 2: W condition



Cluster 3: W condition

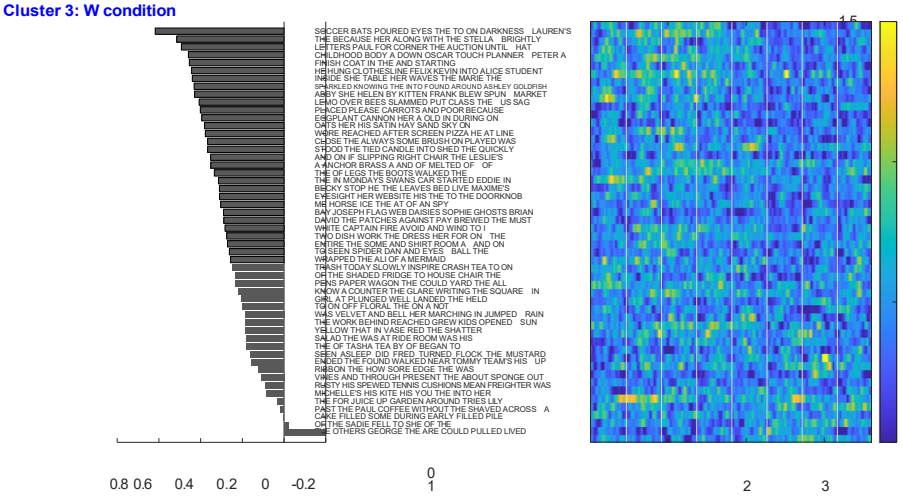


Figure RR3 (added as OSF **Figure O6**). Responses by cluster to individual items from the Word-list condition. The organization of the figure is the same as **Figure RR4**. As with the Sentence condition, individual items within the Word- list condition evoked similar responses.

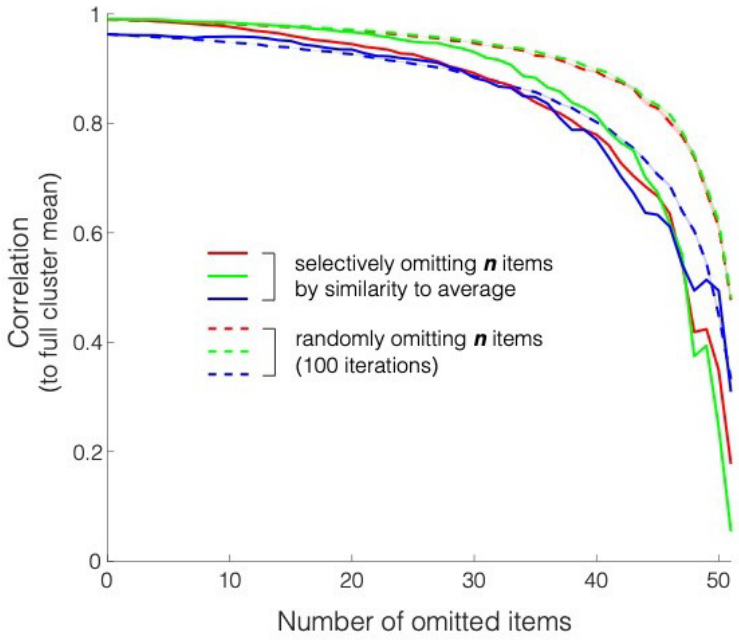


Figure RR4 (added as OSF **Figure O7**). Similarity of average cluster response with progressively more items omitted to average cluster response when all items are used. Items were either (i) removed randomly (dashed lines, 100 random samples of 52-*n* items for a given value of n) or (ii) systemically by their similarity to the condition average (excluding the item being evaluated to avoid information leak, solid lines), where *n* ranged from 0 to 52, the total number of unique items seen by all participants. Similarity to the average cluster response when *no* items were removed was not exactly 1 because there were additional items that a subset of participants (e.g., those with a faster presentation rate) saw that were included in the main analyses (**Figure 2E**), but not here. The systematic removal of items by their similarity to the condition average was intended to reflect the lower bound of similarity: that is, the items being removed that would disrupt the condition average the *most*.

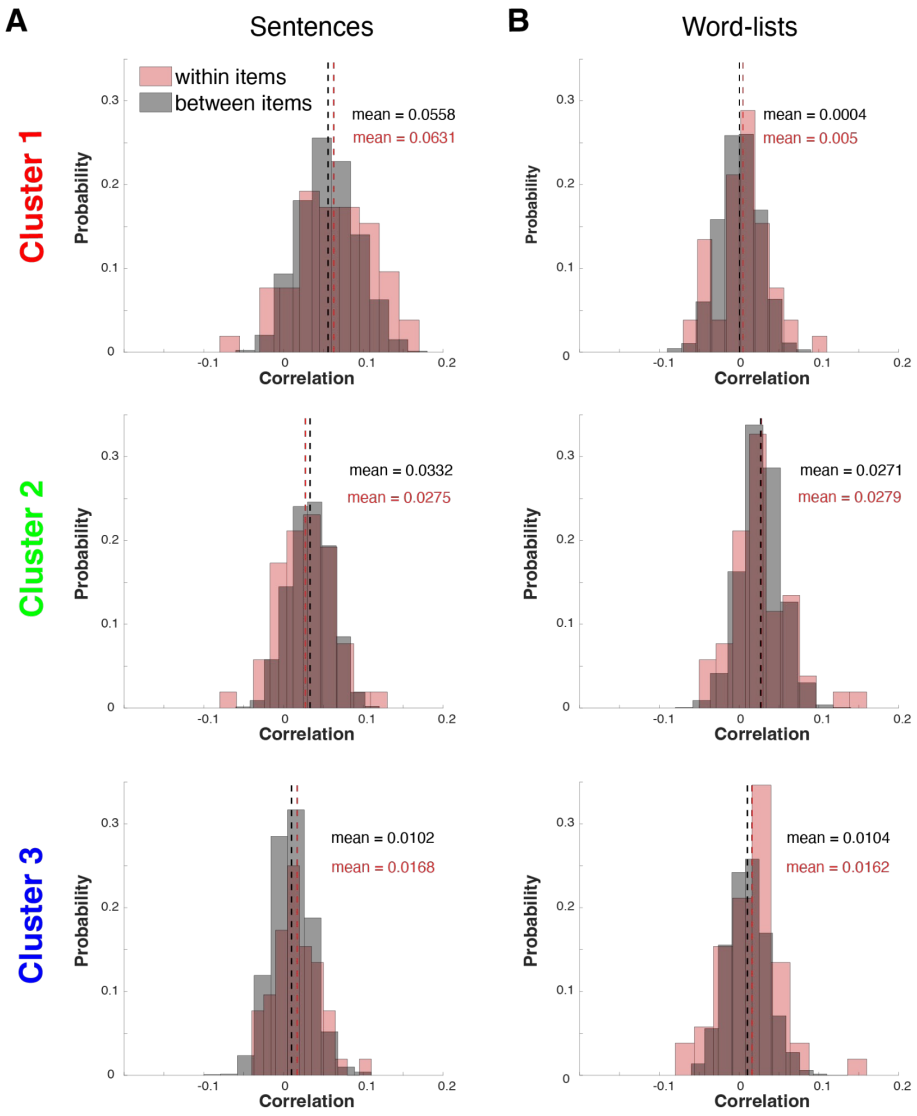


Figure RR5 (updated version of **Figure R8** from the first revision, now **Figure O8** on OSF). Correlation of individual item timecourses across participants within the same item (red) and across different items of the same condition (black) for the Sentence (**A**) and Word-list (**B**) conditions. Item timecourses were constructed by averaging across all electrodes in a particular cluster within a given participant.

Tables

Name	Estimate	SE	tStat	DF	pValue
LME1					
C1, fast - reference	2.61	0.31	8.43	2.17	**0.011**
C2, fast - relative to ref	-1.02	0.27	-3.79	3.36	**0.026**
C3, fast - relative to ref	-2.11	0.50	-4.20	1.75	0.065
C1, slow - relative to ref	2.64	0.48	5.49	4.54	**0.004**
C2, slow - relative to C2, fast	-1.30	0.98	-1.33	2.38	0.296
C3, slow - relative to C3, fast	-1.64	0.88	-1.87	3.76	0.139
LME2					
C1, slow - reference	5.25	0.37	14.35	2.43	**0.002**
C2, slow - relative to ref	-2.32	0.94	-2.47	2.04	0.130
C3, slow - relative to ref	-3.75	0.72	-5.21	2.22	**0.028**
C1, fast - relative to C1, slow	-2.64	0.48	-5.49	4.54	**0.004**
C2, fast - relative to C2, slow	1.30	0.98	1.33	2.38	0.296
C3, fast - relative to C3, slow	1.64	0.88	1.87	3.76	0.139
LME3					
C2, fast - reference	1.60	0.21	7.43	3.05	**0.005**
C3, fast - relative to ref	-1.09	0.44	-2.50	3.82	0.069
C1, fast - relative to ref	1.02	0.27	3.79	3.36	**0.026**
C2, slow - relative to ref	1.34	0.68	1.97	2.46	0.163
C3, slow - relative to C3, fast	-0.34	0.82	-0.42	2.58	0.706
C1, slow - relative to C1, fast	1.30	0.98	1.33	2.38	0.296

LME4					
C3, fast - reference	0.51	0.37	1.39	29.71	0.176
C1, fast - relative to ref	2.11	0.50	4.20	1.75	0.065
C2, fast - relative to ref	1.09	0.44	2.50	3.82	0.069
C3, slow - relative to ref	0.99	0.67	1.48	4.46	0.207
C1, slow - relative to C1, fast	1.64	0.88	1.87	3.76	0.139
C2, slow - relative to C2, fast	0.34	0.82	0.42	2.58	0.706

Table RR1: Fixed Effects results for 4 linear mixed-effects (LME) models: LME1-LME4. All LMEs regress the temporal receptive windows (TRWs) obtained for all electrodes in Dataset 1, on two variables: 1) Cluster (a categorical variable indicating cluster assignment using k-medoid clustering), and 2) Rate (presentation rate of stimuli; three participants had a fast (450ms inter-stimulus interval (ISI)), and 3 had a slow (700 ms ISI) presentation rate). All models used the Wilkinson formula: trw ~ Cluster*Rate + (Cluster*Rate|participant), which corresponds to regressing the TRWs of all electrodes on the fixed effects factors of Cluster and Rate as well as their interaction, adding random effects for all the fixed effects grouped by participant. The variables Cluster and Rate are categorical; Cluster had 3 levels (1, 2, 3) and Rate had 2 levels (slow, fast). The model was coded such that one level from each categorical variable was coded as the reference (intercept, whose estimate was compared to 0 for statistical testing). All other levels of the Cluster variable were modeled relative to the reference, and other levels of Rate were modeled relative to the corresponding estimate (see variable names in table). We ran 4 models that differed in the order of the levels of the categorical variables, such that at each model a different level was coded as the reference. This allowed us to statistically compare all possible pairs of categories, using the LME stats output (Columns 4-6). DF were estimated using the Satterthwaite approximation. Overall all models show a negative trend of TRW by Cluster for both presentation rates (smaller TRWs for C3 relative to C2 and for C2 relative to C1), and an effect of rate only for the TRW of Cluster 1 (larger TRW for C1 with slow relative to fast presentation rates) but no significant effect of rate for Clusters 2 and 3. As a result, the overall main effects of the interaction between Cluster and Rate are not significant due to an additional ANOVA (**Table RR2**).

Name	FStat	DF1	DF2	pValue
LME1				
Intercept (C1, fast)	71.1	1	2.2	**0.01**
Cluster	10.8	2	0	NaN
Rate (C1, slow)	30.2	1	4.5	**0.004**
Cluster:Rate	1.8	2	2.7	0.3
LME2				
Intercept (C1, slow)	205.8	1	2.4	**0.002**
Cluster	14.7	2	0	NaN
Rate (C1, fast)	30.2	1	4.5	**0.004**

Cluster:Rate	1.8	2	2.7	0.321
LME3				
Intercept (C2, fast)	55.2	1	3.1	**0.005**
Cluster	10.8	2	3.5	**0.032**
Rate (C2, slow)	3.9	1	2.5	0.163
Cluster:Rate	1.8	2	2.5	0.329
LME4				
Intercept (C3, fast)	1.9	1	29.7	0.176
Cluster	10.8	2	2.1	0.079
Rate (C3, slow)	2.2	1	4.5	0.207
Cluster:Rate	1.8	2	2.6	0.326

Table RR2: ANOVA results for the LME models above (**Table RR1**). For each model, an F-test is run on 4 effects (first column): 1) Intercept/reference, which is distinct in each specific model, 2) The overall effect of cluster relative to the intercept (collapsing the two levels of the cluster variable that are not the intercept/reference), 3) Rate, which models the effect of the other rate level *above the intercept only*, and 4) The interaction between Cluster and Rate, collapsing all levels beyond the intercept. The interaction between Cluster and Rate is not significant in any model. A p-value of NaN means that there were not sufficient degrees-of-freedom to evaluate the statistic (due to the Satterthwaite approximation).

Decision Letter, second revision:

9th May 2024

Dear Dr. Regev,

Thank you for your patience as we've prepared the guidelines for final submission of your Nature Human Behaviour manuscript, "Neural populations in the language network differ in the size of their temporal receptive windows" (NATHUMBEHAV-23030849B). Please carefully follow the step-by-step instructions provided in the attached file, and add a response in each row of the table to indicate the changes that you have made. Please also address the additional marked-up edits we have proposed within the reporting summary. Ensuring that each point is addressed will help to ensure that your revised manuscript can be swiftly handed over to our production team.

We would hope to receive your revised paper, with all of the requested files and forms within two-three weeks. Please get in contact with us if you anticipate delays.

When you upload your final materials, please include a point-by-point response to any remaining reviewer comments.

If you have not done so already, please alert us to any related manuscripts from your group that are under consideration or in press at other journals, or are being written up for submission to other journals (see: https://www.nature.com/nature-research/editorial-policies/plagiarism#policy-on-duplicate-publication for details).

Nature Human Behaviour offers a Transparent Peer Review option for new original research manuscripts submitted after December 1st, 2019. As part of this initiative, we encourage our authors to support increased transparency into the peer review process by agreeing to have the reviewer comments, author rebuttal letters, and editorial decision letters published as a Supplementary item. When you submit your final files please clearly state in your cover letter whether or not you would like to participate in this initiative. Please note that failure to state your preference will result in delays in accepting your manuscript for publication.

In recognition of the time and expertise our reviewers provide to Nature Human Behaviour's editorial process, we would like to formally acknowledge their contribution to the external peer review of your manuscript entitled "Neural populations in the language network differ in the size of their temporal receptive windows". For those reviewers who give their assent, we will be publishing their names alongside the published article.

Cover suggestions

We welcome submissions of artwork for consideration for our cover. For more information, please see our guide for cover artwork.

If your image is selected, we may also use it on the journal website as a banner image, and may need to make artistic alterations to fit our journal style.

Please submit your suggestions, clearly labeled, along with your final files. We'll be in touch if more information is needed.

ORCID

Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Nature Human Behaviour has now transitioned to a unified Rights Collection system which will allow our Author Services team to quickly and easily collect the rights and permissions required to publish your work. Approximately 10 days after your paper is formally accepted, you will receive an email in providing you with a link to complete the grant of rights. If your paper is eligible for Open Access, our Author Services team will also be in touch regarding any additional information that may be required to arrange payment for your article.

Please note that *Nature Human Behaviour* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and

institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

For information regarding our different publishing models please see our Transformative Journals page. If you have any questions about costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com.

Please use the following link for uploading these materials:
[REDACTED]

If you have any further questions, please feel free to contact me.

Best regards,
[REDACTED]

On behalf of

[REDACTED]

Reviewer #1:
Remarks to the Author:
The authors have given thorough and compelling responses to all concerns raised in the previous round of review. I sincerely appreciate the effort to address even minor concerns with non-trivial supplementary analyses. I also appreciate the minor adjustments to the Discussion re: spatial organization.

I have no other concerns of note.

Reviewer #3:
Remarks to the Author:
I thank the authors for providing further analysis and figures. The additional figures (RR2-4) demonstrate the robustness of the results across trials. Smoothing these figures, e.g., with a 3x3 average filter may make this point more visually explicit, but the authors can decide how best to present this data.

Nima Mesgarani

Final Decision Letter:

Dear Dr Regev,

We are pleased to inform you that your Article "Neural populations in the language network differ in the size of their temporal receptive windows", has now been accepted for publication in Nature Human Behaviour.

Please note that *Nature Human Behaviour* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see http://www.nature.com/nathumbehav/info/gta). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript

submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

We look forward to publishing your paper.

With best regards,
[REDACTED]