High local mutual information drives the response in the human language network

Francis Mollica[1†‡] , Matthew Siegelman[2†], Evgeniia Diachek[2], Steven T. Piantadosi[3], Zachary Mineroff[4], Richard Futrell[2] & Evelina Fedorenko[256‡]

[1]Brain and Cognitive Sciences Department, University of Rochester
[2]Brain and Cognitive Sciences Department/McGovern Institute for Brain Research, MIT
[3]Psychology, University of California, Berkeley
[4]The METALS Program, Carnegie Mellon University
[5]Department of Psychiatry, Harvard Medical School
[6]Department of Psychiatry, Massachusetts General Hospital

[†] F.M.(Author One) and M.S. (Author Two) contributed equally to this work.

[‡] Corresponding authors: mollicaf@gmail.com, evelina9@mit.edu

keywords: language, fMRI, information theory, mutual information, compositionality

Abstract

Traditionally, syntactic operations are thought of as the core computational machinery that sets human language aside from other animal communication systems. Here, we tested an alternative hypothesis: the primary driver of the response in the language-selective regions of the brain is semantic composition. Using formal machinery from information theory, we estimated the likelihood of semantic composition via mutual information among words in a local linguistic context. Across two fMRI experiments, we examined the processing of veridical sentences as well as syntactically degraded sentences, including sentences where the local context does not support semantic composition. Consistent with behavioral/computational modeling work, syntactic degradedness did not lead to lower responses in the fronto-temporal language-selective network, except for when mutual information among words was low. These results challenge the primacy of syntax in the human language architecture, instead supporting the idea that successful semantic composition is what drives the language network in the brain.

A left-lateralized network of frontal and temporal brain regions selectively supports language processing [e.g., 1]. These regions respond to both i) individual word meanings, and ii) combinatorial semantic/syntactic processing [e.g., 2, 3, 4]. The magnitude of neural response appears to scale with how language-like the input is, with the strongest response elicited by sentences, and progressively lower responses elicited by phrases, lists of unconnected words, pseudowords, and foreign or indecipherable speech [5, 6, 7]. But what exact features of the linguistic stimulus drive the language network's response? In particular, sentences—its preferred stimulus—both a) express complex meanings, and b) obey syntactic rules. Is one of these properties dominant?

Many prior proposals have focused on some aspects of syntax [e.g., 6, 8, 9], sometimes with respect to a particular region within the language network, like "Broca's area" [10, 11, 12]. Yet language is fundamentally a vehicle for conveying complex meanings [e.g., 13, 14], and recent work suggests that our sentence interpretation mechanisms are well-designed for coping with syntactic errors ("noise") as long as a plausible meaning is recoverable [e.g., 15, 16, 17, 18]. To evaluate the robustness of the language system to syntactic noise, and inspired by work on lexical processing where word recognition has been shown to be quite resilient to changes in the letter order [19, 20], we used a novel manipulation: naturalistic 12-word-long sentences were gradually degraded by increasing the number of local word swaps (Figure 1), leading to progressively less syntactically well-formed strings (see SI for behavioral evidence of sensitivity to the scrambling manipulation). In an fMRI experiment, 16 participants read these materials (presented using rapid serial visual presentation, one word at a time), along with two control conditions consisting of lists of unconnected words and lists of nonwords. Neural responses were examined in language-responsive regions defined functionally in each individual using a separate localizer task [1], illustrated in Figure 2a-c.

We further hypothesized that the maximal response in the language network would be elicited as long as there is sufficiently high mutual information [21] among the words within a span of a few words, which allows for the construction of complex meanings. At the core of this hypothesis is the idea that the likelihood of an attempt to construct a complex meaning, and probably the ease/success of doing so, is determined by our prior linguistic experience, which, in turn, reflects the distributional properties of objects and events in the world [e.g., 22, 23, 24]. In particular, if we have previously frequently encountered words w_1 and

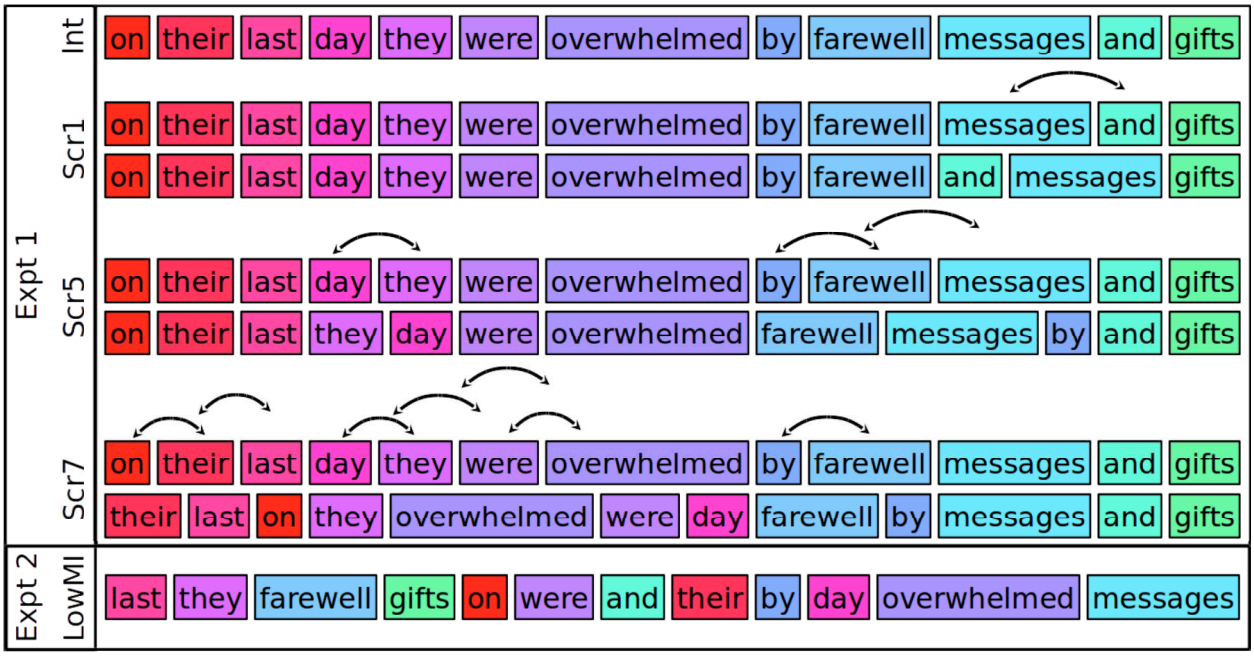

Figure 1: A sample item from the critical experiment; colors are used to illustrate the increasing degradedness (i.e., the color spectrum becomes progressively more discontinuous with more swaps). a. The schematic of the procedure used to create the scrambled-sentence conditions in Experiment 1. b. A sample stimulus from the LowPMI condition in Experiment 2.

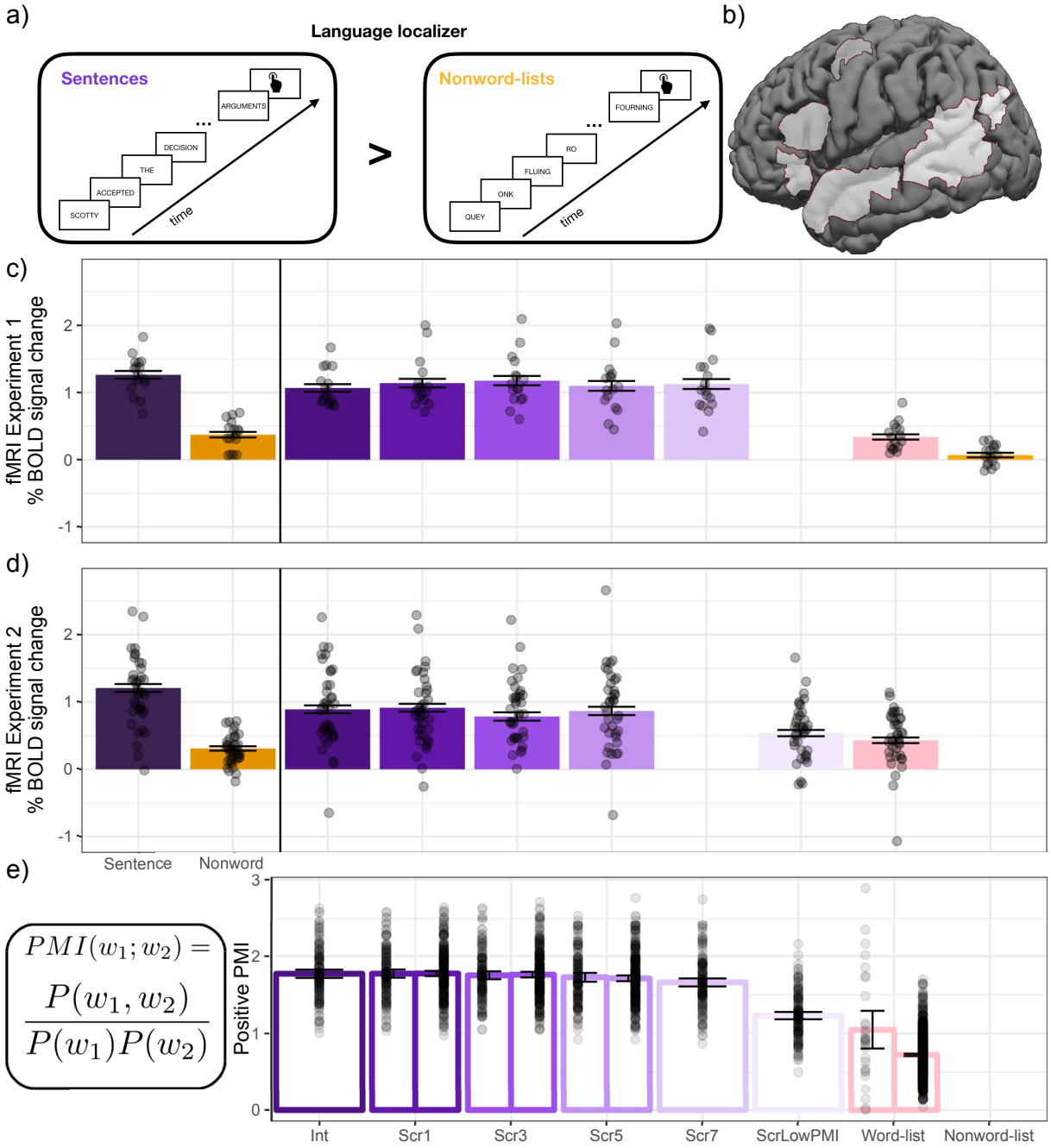

Figure 2: a. The schematic of the language localizer task used to define the language-responsive areas. b. The parcels used to define the language-responsive areas. In each participant, the top 10% of most localizer-responsive voxels within each parcel were taken as that participant's region of interest. Replicating prior work [1], the localizer effect – estimated using across-runs cross-validation to ensure independence – was highly robust in both experiments ($ps < 0.0001$). c-d. Neural responses (in % BOLD signal change relative to fixation) to the conditions of the language localizer and Experiments 1 (n=16) and 2 (n=32). e. The formula for computing pointwise mutual information (PMI) (see Materials and Methods for details), and average positive PMI values for the materials in Experiments 1 and 2 (N.B.: Slightly different scramblings of the materials for the Scr1, Scr3, and Scr5 conditions were used in the two experiments; hence two bars (left=Experiment 1) for each of these conditions.)

w_2 – or words semantically similar to w_1 and w_2—in a semantic dependency, we would be likely to try to combine them again if we encounter them together. For example, if we have frequently encountered words *eat* and *apple*, or *eat* and *restaurant* in a semantic dependency, we will try to combine them if we encounter them again in proximity to one another, even if the syntactic structure within which they occur does not support their semantic composition [15, 17]. If, on the other hand, we have not previously encountered some words together (say *eat* and *accordion*), we may treat them as standalone entities and not attempt to build a complex meaning out of them. The locality constraint is intuitive and presumably arises from our limited working memory capacity [25] and/or our implicit knowledge that words that form semantic dependencies tend to be near each other linearly [26].

To quantify prior linguistic experience, we use pointwise mutual information (PMI), a metric from information theory [27, 28], which measures the mutual dependence between variables (in this case, words). Positive PMI values suggest a semantic dependence between words based on their overlap in contexts of use. Negative and near-zero PMI values suggest the absence of a semantic dependence. Importantly, PMI reflects semantic association, while down-weighting the contribution of high frequency, closed-class words (e.g., *the, it* and *up*). We estimate PMI using a large corpus of English [the Google N-gram corpus; 29]. For each 12-word string, we averaged the positive PMI for all word pairs occurring within a four-word sliding window (see Materials and Methods). We focus on the positive PMI values, because negative values suggest there is no semantic dependency worth building.

To directly evaluate the local mutual information hypothesis, in Experiment 2 (n=32 participants), in addition to some of the same degraded conditions from Experiment 1, we critically included a manipulation where words are scrambled within a sentence in a way that minimizes local PMI and thus, plausibly, the likelihood of complex meaning construction. In particular, we maximized the distance between any two originally proximal content words (Figure 1; see Materials and Methods). The manipulation was effective, leading to a significant drop in local mutual information (Figure 2e). According to the local mutual information hypothesis, the neural response should be substantially lower for this condition relative to the other degraded conditions.

Results

In Experiment 1, replicating much prior work [30, 1, 6], well-formed sentences elicited significantly stronger responses than the word-list and nonword-list conditions (Figure 2c, Table 1). Strikingly, however, degrading the sentences by introducing word swaps did not decrease the magnitude of the language network's response: even stimuli with seven word swaps (e.g., *their last on they overwhelmed were day farewell by messages and gifts*; Figure 1) elicited as strong a response as fully grammatical sentences (e.g., *on their last day they were overwhelmed by farewell messages and gifts*; Figure 2c, Table 1). This pattern of similarly strong neural responses for the well-formed and degraded sentences is in line with the local mutual information hypothesis: mutual information remains high across the sentence conditions (Figure 2e), which plausibly leads to complex meaning construction in spite of syntactic noise.

	Experiment 1				**Experiment 2**			
	Estimate	*Est. Error*	95% *CI*		*Estimate*	*Est. Error*	95% *CI*	
Intact (Int)	1.06†	0.20	0.64	1.46	0.90†	0.18	0.55	1.27
Scrambled (Scr)								
Scr1 vs Int	0.07	0.09	-0.11	0.24	0.03	0.07	-0.11	0.17
Scr3 vs Int	0.11	0.10	-0.08	0.30	-0.10	0.08	-0.25	0.04
Scr5 vs Int	0.03	0.11	-0.19	0.26	-0.02	0.07	-0.16	0.12
Scr7 vs Int	0.06	0.11	-0.15	0.27	–	–	–	–
ScrLowPMI vs Int	–	–	–	–	−0.35†	0.08	-0.52	-0.19
Words vs Int	−0.73†	0.12	-0.97	-0.50	−0.46†	0.08	-0.62	-0.30
Nonwords vs Int	−1.00†	0.13	-1.26	-0.73	–	–	–	–

Table 1: The results of mixed effect linear regressions for Experiments 1 and 2. Condition was dummy-coded with Intact sentences as the reference level. † denotes significant difference.

In Experiment 2, we replicated the pattern observed in Experiment 1 for the intact sentences and sentences with one, three, or five local word swaps, all of which elicited similarly strong responses, all reliably higher than the control word-list condition (Table 1). Critically, in line with the local mutual information hypothesis, the LowPMI condition elicited a neural response that was as low as that elicited by the list of unconnected words (Figure 2b, Table 1).

In spite of eliciting as strong a neural response as veridical sentences, the scrambled sentences are rated as less acceptable behaviorally (SI), suggesting there has to be a cost to the processing of this kind of degraded linguistic input. Indeed, some brain regions outside the fronto-temporal language network—specifically, within the domain-general fronto-parietal multiple demand network [31, 32]—were sensitive to the scrambling manipulation, with stronger responses to more degraded stimuli (SI).

Discussion

Across two fMRI experiments, we found that sufficiently high local mutual information appears to be necessary and sufficient for eliciting the maximal response in the language system, where maximal is defined as the response to the preferred stimulus—well-formed and meaningful sentences. In particular, syntactic degradation—achieved via varying numbers of local word swaps—did not lead to lower neural responses in the frontal or temporal language brain regions, in spite of the fact that such degraded sentences were judged to be less acceptable (SI). Only when there was no evidence in the input for building semantic connections among nearby words did the neural response in the language regions drop to the level of unconnected word lists.

These results suggest that the language system works hard whenever incoming words can be combined into complex meanings [see also 33], in line with current theorizing and evidence from the sentence processing literature whereby our comprehension mechanisms appear to be well-designed for coping with signal corruption [15, 16, 17] and where meaning can drive interpretation even when the syntactic form does not support it [e.g., 34, 35, 36]. However, this result is surprising in light of the emphasis that has traditionally been placed on syntax as the core computational capacity of language [e.g., 37, 38, 39, 10, 12].

Many important questions about semantic composition remain: Is our likelihood of combining any two words affected by our experience with specific lexical items, or would you be as likely to combine *eat* and *sapodilla* even if you only now learned that sapodilla is a type of fruit? What is the span over which high mutual information is detected and affects composition? How is composition influenced by the preceding linguistic and non-linguistic context? For example, if you encounter a string of unrelated words followed by a pair of related words, is the latter sufficient to trigger your semantic composition mechanisms? And is your language system willing to work harder in trying to semantically compose in the presence of top-down knowledge about communicative intent? In spite of all these open questions, a formally specified hypothesis laid out here brings us one step closer to a mechanistic-level account of the computations that the language network plausibly supports.

Methods

Participants

Forty-seven individuals (age $18-48$, average age 22.8; 31 females) participated for payment (Experiment 1: $n = 16$; Experiment 2: $n = 32$; one individual overlapped between the two experiments). (Note that we included twice as many participants in Experiment 2 to ensure that the lack of between-condition differences was not due to insufficient power.) Forty-one were right-handed, as determined by the Edinburgh handedness inventory [40], or by self-report; the remaining six left-handed/ambidextrous individuals showed typical left-lateralized language activations in the language localizer task (see 41 for arguments to include left-handers in cognitive neuroscience experiments). All participants were native speakers of English from the Boston community. Four additional participants were scanned (for Experiment 2) but excluded from the analyses due to excessive head motion or sleepiness, and failure to perform the behavioral task. All participants gave written informed consent in accordance with the requirements of MIT's Committee on the Use of Humans as Experimental Subjects (COUHES).

Experimental Design and Materials

Each participant completed a) a version of the language localizer task [1, Figure 2a], which was used to identify language-responsive areas at the individual subject level, and b) the critical sentence comprehension task (28 participants completed the localizer task in the same session as the critical task, the remaining 19 performed the localizer in an earlier session; see 42 for evidence of the stability of the localizer responses across sessions). Some participants further completed one or two additional tasks for unrelated studies. The entire scanning session lasted approximately 2 hours.

Language localizer

Participants passively read sentences and lists of pronounceable nonwords in a blocked design. The Sentences > Nonwords contrast targets brain regions sensitive to high-level linguistic processing [1]. We have previously established the robustness of this contrast to materials, modality of presentation, language, and task [1, 43, 7]. Each trial started with 100 ms pre-trial fixation, followed by a 12-word-long sentence or a list of 12 nonwords presented on the screen one word/nonword at a time at the rate of 450 ms per word/nonword. Then, a line drawing of a hand pressing a button appeared for 400 ms, and participants were instructed to press a button whenever they saw this icon, and finally a blank screen was shown for 100 ms, for a total trial duration of 6 s. The simple button-pressing task was included to help participants stay awake and focused. Each block consisted of 3 trials and lasted 18 s. Each run consisted of 16 experimental blocks (8 per condition), and five fixation blocks (14 s each), for a total duration of 358 s (5 min 58 s). Each participant performed two runs. Condition order was counterbalanced across runs.

Critical task in Experiment 1

Design and materials. Participants read sentences with correct word order (Intact (Int)) and sentences with progressively more scrambled word orders created by an increasing number of local word swaps (Scrambled (Scr) 1, 3, 5, and 7; Figure 1), as well as two control conditions: lists of unconnected words and nonword lists. At the end of each trial, participants were presented with a word (in the sentence and word-list conditions) or a nonword (in the nonword-list condition) and asked to decide whether this word/nonword appeared in the preceding trial.

To create the sentence materials, we extracted 150 12-word-long sentences from the British National corpus [BNC; 44]. We then permuted the word order in each sentence via local swaps, to create the scrambled conditions. In particular, a word was chosen at random and switched with one of its immediate neighbors. This process was repeated a specified number of times. Because one random swap can directly undo a previous swap, we ensured that the manipulation was successful by calculating the edit distance. (The code used to create the scrambled conditions is available at OSF: https://osf.io/y28fz/) We chose versions with 1, 3, 5, and 7 swaps in order to i) limit the number of sentence conditions to five, while, at the same time, ii) covering a range of degradedness levels. The materials thus consisted of 150 sentences with five versions each (Int, Scr1, Scr3, Scr5, and Scr7), for a total of 750 strings. These were distributed across five experimental lists following a Latin Square design, so that each list contained only one version of a sentence and 30 trials of each of the five conditions. Any given participant saw the materials from just one experimental list, and each list was seen by 2-4 participants. The word-list condition consisted of sequences of 12 real words (173 unique words: 55.5% nouns, 15.6% verbs, 22.5% adjectives, and 6.4% adverbs; average length 7.19 phonemes, standard deviation 1.43; [45]; average log frequency 1.73, standard deviation 0.80 [46]), and the nonword-list condition consisted of sequences of 12 nonwords (there were actually four different nonword-list conditions—a manipulation not of interest to the current study; we averaged the responses across the four nonword-list conditions in the analyses). The word-list and nonword-list materials were the same across participants. (All the materials are available at OSF: https://osf.io/y28fz/)

Procedure. Participants read sentences, scrambled sentences, word lists, and nonword lists in an event-related fMRI design. Each trial lasted 8 s and consisted of the presentation of the stimulus (a sequence of 12 words/nonwords presented one at a time with no punctuation in the center of the screen, for 500 ms each, in black font using capital letters on a white background), followed by a blank screen for 300 ms, followed by a memory probe presented in blue font for 1,200 ms, followed again by a blank screen for 500 ms. The memory probe came from the preceding stimulus on half of the trials. For the sentences, the probes were

uniformly distributed across the beginning (first four words), middle (middle four words), or end (last four words) of the sentence; for the word and nonword lists, the probes were uniformly distributed across the 12 positions. Incorrect probes were the shuffled correct probes from other sequences in the same condition (see SI for behavioral performance).

The trials in each experimental list (300 total; 30 trials per condition, where the conditions included the intact sentence condition, four scrambled sentence conditions, the word-list condition, and four nonword-list conditions) were divided into six subsets corresponding to six runs. Each run lasted 480 s (8 min) and consisted of 8 s * 50 trials (5 per condition) and 80 s of fixation. The optseq2 algorithm [47] was used to create condition orderings and to distribute fixation among the trials so as to optimize our ability to deconvolve neural responses to the different conditions. Condition order varied across runs and participants. Most participants ($n = 13$) performed 5 runs; the remaining participants performed 4 or 3 runs.

Critical task in Experiment 2

Design and materials. As in Experiment 1, participants read sentences with correct word order (Int) and sentences with progressively more scrambled word orders (Scr 1, 3, and 5). The latter three conditions were included in order to assess the robustness and replicability of the results in Experiment 1, in line with recent efforts to increase the reproducibility of findings in cognitive neuroscience [e.g., 48]. (The materials for these scrambled conditions were identical to those in Experiment 1 for half of the participants, and different permutations of the same intact stimuli for the other half.)

The condition with 7 word swaps (Scr7) was replaced by a condition where each pair of nearby content words was separated as much as possible within the 12-word string, so as to minimize local mutual information (Figure 1). We focused on separating nearby content words because those carry the most information in the signal [21]. Take, for example, one of our intact sentences: *Larger firms and international companies tended to offer the biggest pay rises.* First, the content words were given a fixed order that maximized the sum of the distances between adjacent content words (two content words are considered adjacent in the original string if they have no content words between them): e.g., *larger international tended biggest rises firms companies offer pay.* This process was repeated for the function words (e.g., *and the to*). Then, the ordered function words were embedded in the center of the ordered content words (i.e., *larger international tended biggest rises and the to firms companies offer pay*), which maximizes the distances between adjacent content words in the original sentence. (The code is available at OSF: https://osf.io/y28fz/)

In addition to the five sentence conditions, we included five word-list conditions that were matched word-for-word to the sentence conditions. In particular, each of 876 unique words in the sentence conditions was replaced by a different word of the same syntactic category (using the following set: nouns, verbs, adjectives, adverbs, and function words), similar in length (+/- 0.06 phonemes, on average; [45]) and frequency (+/- 0.31 log (lf), on average; [46]). We included the same number of word-list conditions as sentence conditions to match the distribution of sentence and word-/nonword-list conditions in Experiment 1. However, in the analyses, we averaged the responses across the five word-list conditions given that there is no reason to expect a difference among them.

The materials were distributed across five experimental lists; any given participant saw the materials from just one list, and each list was seen by 5-7 participants.

As in Experiment 1, at the end of each trial, participants were presented with a word and asked to decide whether this word appeared in the preceding trial (see SI for behavioral performance).

Procedure. The procedure was identical to that in Experiment 1 except that the memory probe was uniformly distributed across the 12 positions in every condition. Most participants performed 5 or 6 runs (n=30); the remaining participants performed 4 or 3 runs.

fMRI data acquisition. Structural and functional data were collected on the whole-body 3 Tesla Siemens Trio scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted structural images were collected in 179 sagittal slices with 1 mm isotropic voxels (TR = 2,530 ms, TE = 3.48 ms). Functional, blood oxygenation level dependent (BOLD) data were acquired using an EPI sequence (with a 90° flip angle and using GRAPPA with an acceleration factor of 2), with the following acquisition parameters: thirty-one 4 mm thick near-axial slices, acquired in an interleaved order with a 10% distance factor; 2.1 mm x 2.1 mm in-plane resolution; field of view of 200 mm in the phase encoding anterior to posterior (A > P) direction; matrix size of 96 mm

x 96 mm; TR of 2,000 ms; and TE of 30 ms. Prospective acquisition correction [49] was used to adjust the positions of the gradients based on the participant's motion one TR back. The first 10 s of each run were excluded to allow for steady-state magnetization.

fMRI data preprocessing and first-level analysis. First-level analyses were conducted in SPM5 (note that first-level analyses have not changed much in later versions of SPM; we use an older version of the software here due to the use of these data in other projects spanning many years and hundreds of subjects); critical second-level analyses were performed using custom MATLAB and R scripts. Each subject's data were motion corrected and then normalized into a common brain space (the Montreal Neurological Institute (MNI) template) and resampled into 2 mm isotropic voxels. The data were then smoothed with a 4 mm Gaussian filter and high-pass filtered (at 200 s). The task effects in both the language localizer task and the critical experiment were estimated using a General Linear Model (GLM) in which each experimental condition was modeled with a boxcar function (corresponding to a block or event) convolved with the canonical hemodynamic response function (HRF).

Language fROI definition and response estimation. For each participant, functional regions of interest (fROIs) were defined using the Group-constrained Subject-Specific (GSS) approach [1], whereby a set of parcels or "search spaces" (i.e., brain areas within which most individuals in prior studies showed activity for the localizer contrast) is combined with each individual participant's activation map for the same contrast. To define the language fROIs, we used six parcels (Figure 2b) derived from a group-level representation of data for the Sentences>Nonwords contrast in 220 participants. These parcels included three regions in the left frontal cortex: two located in the inferior frontal gyrus (LIFG and LIFGorb), and one located in the middle frontal gyrus (LMFG); and three regions in the left temporal and parietal cortices spanning the entire extent of the lateral temporal lobe and extending into the angular gyrus (LAntTemp, LPostTemp, and LAngG). (These parcels were similar to the parcels reported originally in [1], except that the two anterior temporal regions got grouped together, and the two posterior temporal regions got grouped together.) Following much prior work in our group, individual fROIs were defined by selecting—within each parcel—the top 10% of most localizer-responsive voxels based on the t values for the Sentences>Nonwords contrast. Responses (in percent BOLD signal change units) to the relevant critical experiment's conditions, relative to the fixation baseline, were then extracted from these fROIs. So, the input to the critical statistical analyses consisted of—for each participant—a value (percent BOLD signal change) for each of 10 conditions in each of the six language fROIs. Further, for Experiment 1, responses were averaged across the four nonword-list conditions, leaving a total of seven conditions; and for Experiment 2, responses were averaged across the five word-list conditions, leaving a total of six conditions. In all the critical analyses, we consider the language network as a whole (treating regions as random effects; see below) given the abundant evidence that the regions of this network form an anatomically [e.g., 50] and functionally integrated system, as evidenced by strong inter-regional correlations during rest and language comprehension [e.g., 51]), but see Figure S1 and Table S3 for the six language fROIs' profiles and associated statistics.

Computing mutual information values. We used a sliding four-word window to extract local word pairs from each 12-word string. This is equivalent to collecting the bigrams, 1-skip-grams and 2-skip-grams from each string. For each word pair, we calculated PMI as follows:

$$PMI(w_i; w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \tag{1}$$

Probabilities were estimated using the Google N-gram corpus [29] and ZS [52] with Laplace smoothing ($\alpha = 0.1$). For each 12-word string, we averaged across the positive PMI values. (The code for computing PMI is available at OSF: https://osf.io/y28fz/)

Statistical tests. To compare the average change in BOLD response across conditions, we conducted a mixed effect linear regression model with maximal random effect structure, predicting the level of BOLD response with a fixed effect and random slopes for Condition, and random effects for Region of Interest and Participant. Condition was dummy-coded with Intact sentences as the reference level. Models were fit separately for Experiment 1 and Experiment 2 using the brms package [53] in R [54] to interface with Stan [55]. Results are presented in Table 1. (Data and analysis code are available at OSF: https://osf.io/y28fz/)

Acknowledgements

We thank i) Zuzanna Balewski for help with creating the experimental script for Experiment 1, ii) EvLab members for help with scanning and helpful discussions, and iii) Nancy Kanwisher, Ted Gibson, Adele Goldberg, Leon Bergen, Josh Tenenbaum, and the audience at the CUNY2017 Sentence Processing conference for comments on this line of work. The authors would also like to acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, and the support team (Steven Shannon, Atsushi Takahashi, and Sheeba Arnold). E.F. was supported by NIH awards R00-HD-057522 and R01-DC-016607.

Competing Interests

We declare no competing interests.

References

[1] Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194, 2010.

[2] Timothy A Keller, Patricia A Carpenter, and Marcel Adam Just. The neural bases of sentence comprehension: a fmri examination of syntactic and lexical processing. *Cerebral cortex*, 11(3):223–237, 2001.

[3] Evelina Fedorenko, Alfonso Nieto-Castanon, and Nancy Kanwisher. Lexical and syntactic representations in the brain: an fmri investigation with multi-voxel pattern analyses. *Neuropsychologia*, 50(4): 499–513, 2012.

[4] Alexa Bautista and Stephen M Wilson. Neural responses to grammatically and lexically degraded speech. *Language, cognition and neuroscience*, 31(4):567–574, 2016.

[5] Marina Bedny, Alvaro Pascual-Leone, David Dodell-Feder, Evelina Fedorenko, and Rebecca Saxe. Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, 108(11):4429–4434, 2011.

[6] Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, page 201018711, 2011.

[7] Terri L Scott, Jeanne Gallée, and Evelina Fedorenko. A new fun and robust version of an fmri localizer for the frontotemporal language system. *Cognitive neuroscience*, 8(3):167–176, 2017.

[8] Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1):158, 2016.

[9] Matthew J Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S Cash, Lionel Naccache, John T Hale, Christophe Pallier, et al. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678, 2017.

[10] Angela D Friederici, Jörg Bahlmann, Stefan Heim, Ricarda I Schubotz, and Alfred Anwander. The brain differentiates human and non-human grammars: functional localization and structural connectivity. *Proceedings of the National Academy of Sciences*, 103(7):2458–2463, 2006.

[11] Lorraine K Tyler, William D Marslen-Wilson, Billi Randall, Paul Wright, Barry J Devereux, Jie Zhuang, Marina Papoutsi, and Emmanuel A Stamatakis. Left inferior frontal cortex and syntax: function, structure and behaviour in patients with left hemisphere damage. *Brain*, 134(2):415–431, 2011.

[12] Robert C Berwick, Angela D Friederici, Noam Chomsky, and Johan J Bolhuis. Evolution, brain, and the nature of language. *Trends in cognitive sciences*, 17(2):89–98, 2013.

[13] Adele E Goldberg. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand, 2006.

[14] Ray Jackendoff. What is the human language faculty? two views. *Language*, pages 586–624, 2011.

[15] Fernanda Ferreira, Karl GD Bailey, and Vittoria Ferraro. Good-enough representations in language comprehension. *Current directions in psychological science*, 11(1):11–15, 2002.

[16] Roger Levy, Klinton Bicknell, Tim Slattery, and Keith Rayner. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090, 2009.

[17] Edward Gibson, Leon Bergen, and Steven T Piantadosi. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, page 201216438, 2013.

[18] Matthew J Traxler. Trends in syntactic parsing: Anticipation, bayesian estimation, and good-enough parsing. *Trends in cognitive sciences*, 18(11):605–611, 2014.

[19] Jonathan Grainger and Walter JB Van Heuven. Modeling letter position coding in printed word perception. In P. Bonin, editor, *Mental lexicon: "Some words to talk about words"*, pages 1–23. Nova Science Publishers, Hauppauge, NY, US, 2004.

[20] Colin J Davis. The spatial coding model of visual word identification. *Psychological review*, 117(3):713, 2010.

[21] Claude E Shannon and Warren Weaver. The mathematical theory of communication. 1949. *Urbana, IL: University of Illinois Press*, 1963.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[24] Francisco Pereira, Samuel Gershman, Samuel Ritter, and Matthew Botvinick. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive neuropsychology*, 33(3-4):175–190, 2016.

[25] Edward Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76, 1998.

[26] Richard Landy Jones Futrell. *Memory and locality in natural language*. PhD thesis, Massachusetts Institute of Technology, 2017.

[27] R. M. Fano. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, 1961.

[28] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[29] Jean-Baptiste Michel, Yuan Kui Shen, Aviva P Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, page 1199644, 2010.

[30] Tineke M Snijders, Theo Vosse, Gerard Kempen, Jos JA Van Berkum, Karl Magnus Petersson, and Peter Hagoort. Retrieval and unification of syntactic structure in sentence comprehension: an fmri study using word-category ambiguity. *Cerebral cortex*, 19(7):1493–1503, 2008.

[31] John Duncan. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, 14(4):172–179, 2010.

[32] John Duncan. The structure of cognition: attentional episodes in mind and brain. *Neuron*, 80(1):35–50, 2013.

[33] Evelina Fedorenko, Terri L Scott, Peter Brunner, William G Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41):E6256–E6262, 2016.

[34] Whitney Tabor, Bruno Galantucci, and Daniel Richardson. Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4):355–370, 2004.

[35] Albert Kim and Lee Osterhout. The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of memory and language*, 52(2):205–225, 2005.

[36] Gina R Kuperberg, David Caplan, Tatiana Sitnikova, Marianna Eddy, and Phillip J Holcomb. Neural correlates of processing syntactic, semantic, and thematic relationships in sentences. *Language and cognitive processes*, 21(5):489–530, 2006.

[37] Noarn Chomsky. Language and mind. harcourt brace javanovich. *New York*, 1972.

[38] Steven Pinker. The language instinct: The new science of language and mind (vol. 7529). *UK: Penguin*, 1995.

[39] Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

[40] Richard C Oldfield. The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, 9(1):97–113, 1971.

[41] Roel M Willems, Lise Van der Haegen, Simon E Fisher, and Clyde Francks. On the other hand: including left-handers in cognitive neuroscience and neurogenetics. *Nature Reviews Neuroscience*, 15(3):193, 2014.

[42] Kyle Mahowald and Evelina Fedorenko. Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *Neuroimage*, 139:74–93, 2016.

[43] Evelina Fedorenko. The role of domain-general cognitive control in language comprehension. *Frontiers in psychology*, 5:335, 2014.

[44] Lou Burnard. Reference guide for the british national corpus (world edition), 2000.

[45] Robert Weide. The cmu pronunciation dictionary, release 0.6, 1998.

[46] Marc Brysbaert, Boris New, and Emmanuel Keuleers. Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, 44(4):991–997, 2012.

[47] Anders M Dale. Optimal experimental design for event-related fmri. *Human brain mapping*, 8(2-3): 109–114, 1999.

[48] Russell A Poldrack, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115, 2017.

[49] Stefan Thesen, Oliver Heid, Edgar Mueller, and Lothar R Schad. Prospective acquisition correction for head motion with image-based tracking for real-time fmri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 44(3):457–465, 2000.

[50] Dorothee Saur, Björn W Kreher, Susanne Schnell, Dorothee Kümmerer, Philipp Kellmeyer, Magnus-Sebastian Vry, Roza Umarova, Mariacristina Musso, Volkmar Glauche, Stefanie Abel, et al. Ventral and dorsal pathways for language. *Proceedings of the national academy of Sciences*, pages pnas–0805234105, 2008.

[51] Idan Blank, Nancy Kanwisher, and Evelina Fedorenko. A functional dissociation between language and multiple-demand systems revealed in patterns of bold signal fluctuations. *Journal of neurophysiology*, 112(5):1105–1118, 2014.

[52] Nathaniel J Smith. Zs: A file format for efficiently distributing, using, and archiving record-oriented data sets of any size.

[53] Paul-Christian Bürkner et al. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28, 2017.

[54] R Core Team. R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria, 2017.

[55] Stan Development Team. RStan: the R interface to Stan, 2018. URL `http://mc-stan.org/`. R package version 2.17.3.

[56] Alfonso Nieto-Castañón and Evelina Fedorenko. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *Neuroimage*, 63(3):1646–1669, 2012.

[57] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535, 2009.

[58] Evelina Fedorenko, John Duncan, and Nancy Kanwisher. Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, page 201315235, 2013.

[59] John Duncan and Adrian M Owen. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in neurosciences*, 23(10):475–483, 2000.

[60] Kenneth Hugdahl, Marcus E Raichle, Anish Mitra, and Karsten Specht. On the existence of a generalized non-specific task-dependent network. *Frontiers in human neuroscience*, 9:430, 2015.

[61] Joseph H Callicott, Venkata S Mattay, Alessandro Bertolino, Kimberly Finn, Richard Coppola, Joseph A Frank, Terry E Goldberg, and Daniel R Weinberger. Physiological characteristics of capacity constraints in working memory as revealed by functional mri. *Cerebral cortex*, 9(1):20–26, 1999.

[62] David EJ Linden, Robert A Bittner, Lars Muckli, James A Waltz, Nikolaus Kriegeskorte, Rainer Goebel, Wolf Singer, and Matthias HJ Munk. Cortical capacity constraints for visual working memory: dissociation of fmri load effects in a fronto-parietal network. *Neuroimage*, 20(3):1518–1530, 2003.

[63] Tanya Wen, Daniel J Mitchell, and John Duncan. Response of the multiple-demand network during simple stimulus discriminations. *NeuroImage*, 177:79–87, 2018.

Supplementary Materials

Acceptability ratings for the materials used in Experiment 1.

To ensure that our scrambling manipulation was successful, 76 participants recruited through Amazon.com's Mechanical Turk rated the naturalness of the stimuli used in Experiment 1 on a 7-point scale (from 1=unnatural to 7=natural). The ratings were analyzed using a mixed effect linear regression model with a fixed effect and random slopes for Condition, and random effects for Participant and Item. To demonstrate the effectiveness of the manipulation at every level, Condition was backwards difference coded. As can be seen in Table S1, every increase in degradation was rated as significantly less natural than the previous amount of degradation, although with diminishing returns. Thus participants were robustly sensitive to the degradation manipulation. Code and data are available at OSF: https://osf.io/y28fz/.

	Estimate	*Est. Error*	95% *CI*	
Grand Mean	3.54†	0.08	3.38	3.69
Scr1 - Int	−2.04†	0.11	-2.25	-1.83
Scr3 - Scr1	−1.42†	0.08	-1.57	-1.26
Scr5 - Scr3	−0.56†	0.05	-0.67	-0.46
Scr7 - Scr5	−0.23†	0.04	-0.30	-0.15

Table S1: The results of a mixed effect linear regression for the acceptability rating data. † denotes significant difference.

Behavioral (memory probe task) data in Experiments 1 and 2.

To ensure that participants were attentive across conditions, we included a memory probe task in both Experiments 1 and 2. After each stimulus (sentence, word list, or nonword list), participants saw a probe word/nonword and were asked to press one of two buttons to indicate whether this word/nonword appeared in the string they just read. Response accuracy for each experiment was analyzed with a mixed effect logistic regression model with a fixed effect and random slopes for Condition, and random intercepts for Participant and Item. Condition was dummy-coded with Intact Sentences as the reference level. For both experiments, accuracy was above chance for all conditions. In Experiment 1, accuracies in the scrambled sentence conditions did not significantly differ from accuracy in the intact sentence condition; however, accuracy was significantly lower in the word-list and nonword-list conditions compared to the intact sentence condition (Table S2), in line with prior work [e.g., 1]. Similarly, in Experiment 2, accuracies in the scrambled sentence conditions did not significantly differ from accuracy in the intact sentence condition; however, accuracy was lower in the lowPMI and the word-list conditions compared to the intact sentence condition (Table S2). Data and analysis code are available at OSF: https://osf.io/y28fz/.

	Experiment 1				Experiment 2			
	Estimate	*Est. Error*	95% *CI*					
Intact (Int)	3.63†	0.33	3.03	4.32	1.98†	0.24	1.52	2.46
Scrambled (Scr)								
Scr1 vs Int	0.37	0.86	-0.94	2.44	-0.19	0.22	-0.62	0.27
Scr3 vs Int	-0.02	0.74	-1.19	1.76	0.11	0.27	-0.39	0.70
Scr5 vs Int	-0.39	0.59	-1.39	0.94	-0.27	0.24	-0.71	0.22
Scr7 vs Int	-0.45	0.61	-1.50	0.87	–	–	–	–
ScrLowPMI vs Int	–	–	–	–	−0.64†	0.23	-1.07	-0.16
Words vs Int	−1.74†	0.38	-2.51	-1.01	−0.30†	0.17	-0.64	0.04
Nonwords vs Int	−2.35†	0.33	-3.04	-1.75	–	–	–	–

Table S2: The results of logistic mixed effect models for Experiments 1 and 2 for the memory probe data. Stimulus Type was dummy-coded with Intact sentences as the reference level. † denotes significant difference.

Individual language regions' responses to the conditions of Experiment 1 and 2.

In the main analysis (Figure 2c-d, Table 1), we reported a model that examined the response in the language network as a whole (treating the six regions as random effects). As shown in Figure S1 and Table S3, the pattern observed across the network was present—both qualitatively and statistically—in each of the six language fROIs individually.

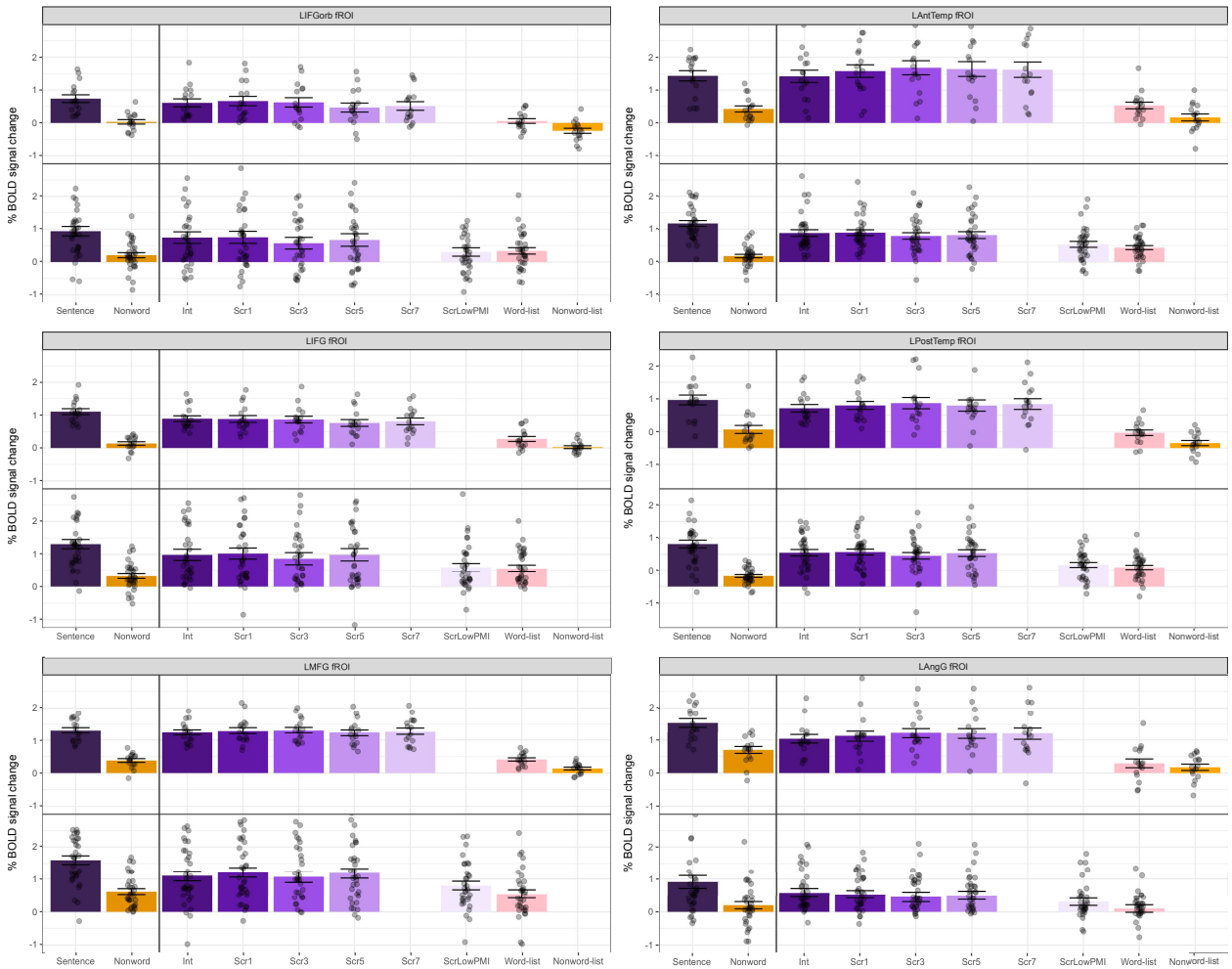

Figure S1: Neural responses (in % BOLD signal change relative to fixation) to the conditions of the language localizer and Experiments 1 (top panel) and 2 (bottom panel) in each of the six language fROIs.

	Experiment 1				Experiment 2			
	Estimate	*Est. Error*	95% *CI*		*Estimate*	*Est. Error*	95% *CI*	
Intact (Int)								
LIFGorb	0.61†	0.13	0.35	0.85	0.75†	0.18	0.40	1.10
LIFG	0.87†	0.09	0.71	1.07	0.98†	0.17	0.64	1.31
LMFG	1.24†	0.08	1.07	1.41	1.11†	0.16	0.81	1.42
LAntTemp	1.41†	0.19	1.03	1.80	0.89†	0.10	0.70	1.09
LPostTemp	1.20†	0.12	0.97	1.44	1.04†	0.10	0.85	1.24
LAngG	1.04†	0.15	0.75	1.33	0.60†	0.14	0.33	0.86
Scr1 vs Int								
LIFGorb	0.06	0.05	-0.05	0.16	0.01	0.09	-0.17	0.19
LIFG	-0.01	0.05	-0.10	0.09	0.04	0.07	-0.09	0.17
LMFG	0.05	0.06	-0.07	0.17	0.11†	0.06	0.00	0.23
LAntTemp	0.16	0.10	-0.03	0.36	0.01	0.05	-0.10	0.12
LPostTemp	0.08	0.09	-0.10	0.27	0.02	0.06	-0.09	0.13
LAngG	0.09	0.07	-0.05	0.21	-0.05	0.06	-0.17	0.07
Scr3 vs Int								
LIFGorb	0.02	0.06	-0.10	0.12	-0.17	0.09	-0.36	0.01
LIFG	-0.02	0.07	-0.17	0.11	-0.12	0.08	-0.28	0.03
LMFG	0.07	0.07	-0.07	0.19	-0.03	0.07	-0.17	0.12
LAntTemp	0.26†	0.11	0.05	0.47	-0.09	0.06	-0.20	0.02
LPostTemp	0.15	0.12	-0.08	0.40	-0.09	0.06	-0.21	0.03
LAngG	0.18†	0.08	0.03	0.33	-0.13	0.08	-0.28	0.03
Scr5 vs Int								
LIFGorb	−0.14†	0.07	-0.28	-0.01	-0.07	0.10	-0.26	0.12
LIFG	−0.13†	0.05	-0.22	-0.04	0.00	0.07	-0.14	0.14
LMFG	-0.02	0.08	-0.19	0.13	0.08	0.07	-0.06	0.21
LAntTemp	0.22†	0.12	0.00	0.46	-0.07	0.06	-0.18	0.04
LPostTemp	0.08	0.12	-0.16	0.31	-0.02	0.06	-0.14	0.10
LAngG	0.16†	0.08	0.01	0.32	-0.08	0.06	-0.21	0.04
Scr7 vs Int								
LIFGorb	-0.09	0.07	-0.22	0.04	–	–	–	–
LIFG	-0.08	0.06	-0.21	0.08	–	–	–	–
LMFG	0.03	0.08	-0.14	0.19	–	–	–	–
LAntTemp	0.21	0.16	-0.10	0.50	–	–	–	–
LPostTemp	0.13	0.13	-0.13	0.38	–	–	–	–
LAngG	0.16	0.11	-0.05	0.37	–	–	–	–
ScrLowPMI vs Int								
LIFGorb	–	–	–	–	−0.44†	0.11	-0.66	-0.22
LIFG	–	–	–	–	−0.39†	0.09	-0.57	-0.21
LMFG	–	–	–	–	−0.29†	0.08	-0.46	-0.13
LAntTemp	–	–	–	–	−0.34†	0.06	-0.46	-0.22
LPostTemp	–	–	–	–	−0.37†	0.06	-0.50	-0.25
LAngG	–	–	–	–	−0.28†	0.08	-0.43	-0.13
Words vs Int								
LIFGorb	−0.55†	0.09	-0.72	-0.37	−0.41†	0.13	-0.67	-0.16
LIFG	−0.62†	0.07	-0.76	-0.48	−0.43†	0.10	-0.62	-0.23
LMFG	−0.83†	0.07	-0.98	-0.68	−0.55†	0.07	-0.68	-0.40
LAntTemp	−0.88†	0.14	-1.14	-0.61	−0.44†	0.07	-0.58	-0.31
LPostTemp	−0.74†	0.11	-0.96	-0.54	−0.45†	0.07	-0.58	-0.31
LAngG	−0.75†	0.13	-1.00	-0.48	−0.49†	0.07	-0.63	-0.36
Nonwords vs Int								
LIFGorb	−0.84†	0.13	-1.09	-0.57	–	–	–	–
LIFG	−0.86†	0.06	-0.99	-0.74	–	–	–	–
LMFG	−1.11†	0.08	-1.27	-0.95	–	–	–	–
LAntTemp	−1.24†	0.15	-1.55	-0.94	–	–	–	–
LPostTemp	−1.06†	0.12	-1.29	-0.83	–	–	–	–
LAngG	−0.87†	0.11	-1.07	-0.65	–	–	–	–

Table S3: The results of mixed effect linear regressions for Experiments 1 and 2 for the six language fROIs. Condition was dummy-coded with Intact sentences as the reference level. † denotes significant difference.

Brain regions sensitive to the sentence-scrambling manipulation.

A critical result observed in both Experiments 1 and 2 is the lack of neural response reduction in the language regions for sentences where word order has been permuted yet the level of PMI among nearby words remained high (which, we hypothesize, allowed for semantic composition). Behaviorally (see above), we found that participants are highly sensitive to the scrambling manipulation as evidenced by progressively lower acceptability ratings for more scrambled sentences. Here, we asked whether any parts of our brain work harder when we process more scrambled sentences.

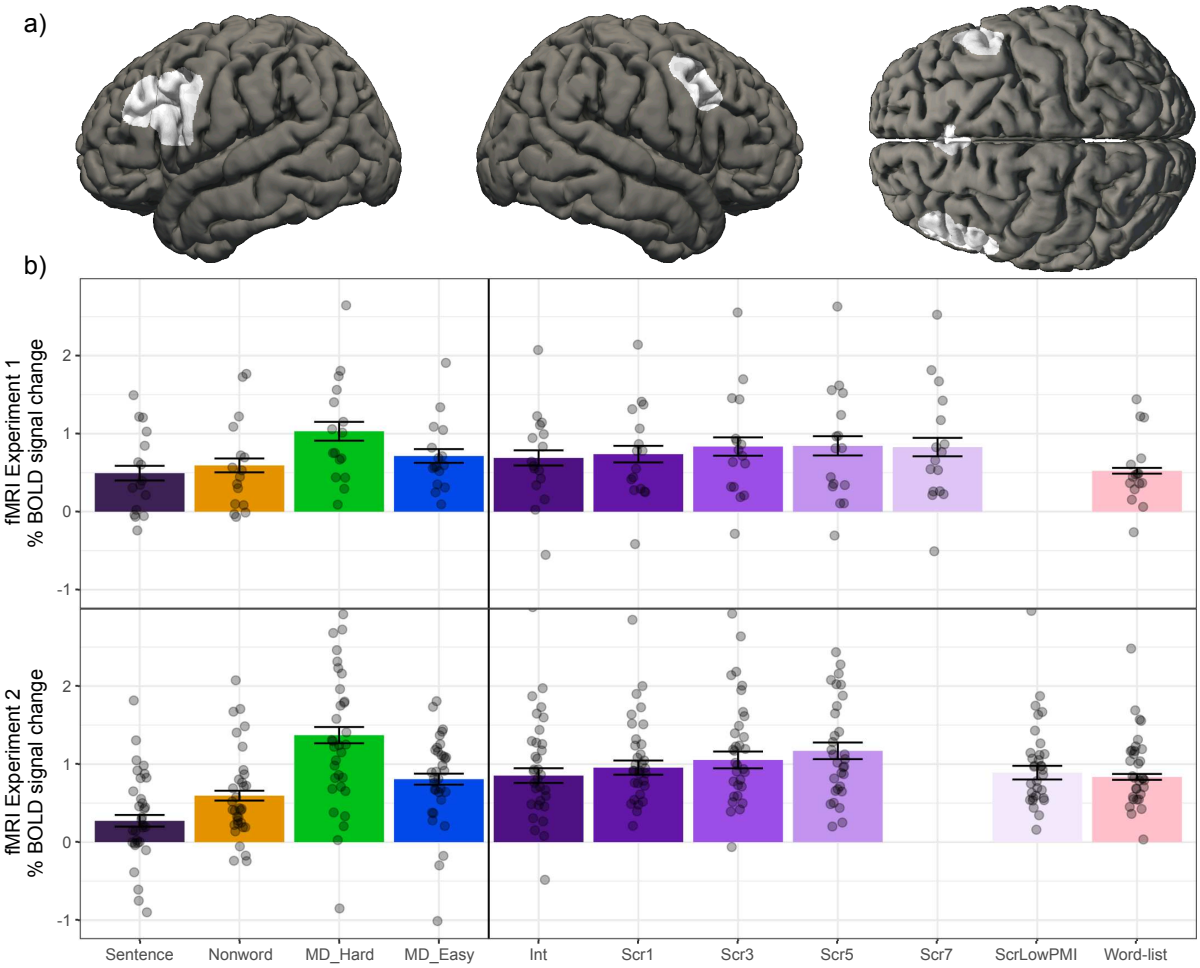

Figure S2: a. The parcels used to define the scrambling-responsive areas. In each participant, the top 10% of most localizer-responsive voxels within each parcel were taken as that participant's region of interest. b. Neural responses (in % BOLD signal change relative to fixation) to the conditions of Experiments 1 (top) and 2 (bottom), as well as the language localizer and spatial WM task.

Discovery of scrambling-responsive regions. To search for brain regions sensitive to scrambling, we performed a group-constrained subject-specific (GSS) whole-brain analysis [1]. This analysis searches for spatially consistent (across individuals) patterns of activation while taking into account inter-individual variability in the precise loci of activations, which increases sensitivity relative to traditional random-effects analyses that assume voxel-wise correspondence across people [56]. We chose a contrast between the most scrambled condition that was shared between the two experiments (i.e., Scr5) and the intact condition. Pooling data across experiments (n=47), we took individual whole-brain activation maps for the Scr5>Int contrast and binarized them so that voxels that show a reliable effect (significant at $p < 0.05$, uncorrected at the whole-brain level) were turned into 1's and all other voxels were turned into 0's. (We chose a liberal

	Estimate	*Est. Error*	95% *CI*	
Intercept	0.844	0.491	-0.125	1.835
Edit Distance	0.046†	0.024	0.012	0.086
Edit Distance^2	−0.002†	0.001	-0.003	-0.0005

Table S4: The results of mixed effect linear regression models for the scrambling-responsive regions. † denotes significant difference.

threshold for the individual activation maps to maximize our chances of detecting regions of interest; as explained below, however, the resulting regions were subsequently evaluated using statistically conservative criteria.) We overlaid these maps to create a probabilistic activation overlap map, thresholded this map to only include voxels where at least 4 of the 47 participants showed activation, and divided it into "parcels" using a watershed image parcellation algorithm (see [1] for details). Finally, we identified parcels that—when intersected with the individual activation maps—contained supra-threshold (i.e., significant for our contrast of interest at $p < 0.05$, uncorrected) voxels in at least half of the individual participants. Four parcels satisfied this criterion, located in the middle frontal gyrus bilaterally and in the SMA (Figure S2a).

Characterizing the functional profiles of scrambling-responsive regions in greater detail. In each of the four regions, we estimated the response magnitude to the conditions of the two experiments. To estimate the responses to the Scr5 and Int conditions, which were used in the localizer contrast, we used an across-runs cross-validation procedure [e.g., 56], to ensure independence between the data used to define the fROIs and to estimate the responses [57]. In particular, each parcel was intersected with each participant's activation map for the Scr5>Int contrast for all but the first run of the data. The voxels within the parcel were sorted—for each participant—based on their t-values, and the top 10% of voxels were selected as that participant's fROI. The responses were then estimated using the left out run's data. The procedure was repeated iteratively leaving out each of the runs. Finally, the responses were averaged across the left-out runs to derive a single response magnitude per subject per region per condition. To estimate the responses to the other critical conditions, we used data from the Scr5 and Int conditions across all runs. Statistical tests were performed on these extracted percent BOLD signal change values.

To more richly characterize the functional profiles of these regions, we estimated the responses to two other experiments: i) the language localizer, and ii) a spatial working memory (WM) experiment, which we often include in our studies for control analyses. Responses to the language localizer conditions can tell us whether the scrambling-responsive regions show a signature of the language network: i.e., stronger responses to sentences than nonword sequences. We have constrained our definition of the language-responsive regions in the critical analyses by a set of parcels derived based on activations for the language localizer contrast in a large number of individuals. Thus regions outside of this network of language-responsive regions should not show language-responsive properties. So this analysis provides a reality check of sorts.

Responses to the conditions of the spatial WM task tell us whether the scrambling-responsive regions belong to the domain-general multiple demand (MD) network, which has been implicated in executive functions like working memory and cognitive control [31, 32]. In this task, participants have to keep track of four (easy condition) or eight (hard condition) locations in a 3 x 4 grid, and then perform a two-alternative, forced-choice task at the end of each trial to indicate the set of locations that they just saw. The Hard>Easy contrast has been shown to robustly and reliably activate the regions of the MD network [58, 51]. The timing and procedure details can be found in [58], but briefly: each block consisted of four 8 s trials and lasted 32 s, and each run consisted of 12 experimental blocks (6 per condition), and 4 fixation blocks (16 s in duration each), for a total duration of 448 s (7 min 28 s). 45 participants performed two runs of the task; the remaining 2 participants performed one run of the task. Condition order was counterbalanced across runs when participants performed two runs.

The patterns of responses observed are shown in Figures S2b (across the fROIs) and S3 (for the four fROIs individually). Qualitatively, with respect to the conditions of the critical experiments, we found that the response increases parametrically from the Intact to the Scr5 condition in both experiments. Further, in Experiment 1, the response stays high or continues to increase for the Scr7 condition, but in Experiment 2, the response appears to fall off for the LowPMI condition. To quantify this non-monotonic pattern, we collapsed across experiments and conducted a mixed effect linear regression with first and second order terms

for Edit Distance[1] (i.e., the number of swaps required to reconstruct the original intact sentence) as fixed effects and random slopes, and random effects for Participant and Region of Interest. We find a significant initial increase in response which decreases slowly as stimuli become more degraded (Table S4).

With respect to the conditions of the language localizer and the spatial WM experiment, we found that none of the four fROIs showed a stronger response to sentences than nonword sequences (in fact, three of the four regions showed a reliably stronger response to nonword sequences than sentences); and that all four fROIs showed a stronger response to the Hard than Easy WM condition. These results suggest that the scrambling-responsive fROIs fall within the domain-general MD cortex [31, 32]. The parametric increase as a function of the degree of scrambling in the critical experiments is in line with robust sensitivity of the MD cortex to effort across domains [e.g., 59, 58, 60]. In particular, participants have to work harder to extract meaning from the more scrambled sentences. The fall-off for the LowPMI condition – which elicited a low response in the language regions as shown in our critical analysis—fits with the idea that participants give up their attempts to derive a meaningful representation in his condition (e.g., 61, 62 cf. 63).

In summary, we found a number of regions that fall within the domain-general multiple demand network [31, 32] that work harder when participants process sentences with scrambled word orders compared to intact sentences. The level of response increases as the degree of scrambling increases, until participants are no longer able to construct a complex meaning (as evidenced by a low response in the language network), which occurs in the LowPMI condition. The increase as a function of scrambling is in stark contrast to the pattern we observed in the language-responsive regions, which responded similarly strongly to intact and scrambled sentences. These results suggest that participants do experience greater cost for processing scrambled sentences but this cost is carried by domain-general regions that support diverse demanding tasks across domains. The language regions instead respond similarly strongly as long as local mutual information remains high, which we argue allows for complex meaning construction.

[1]Switching from a categorical (Condition) to continuous coding (Edit Distance) permits testing non-monotonicity and increases the sensitivity to detect an effect. To ensure that our analysis of the language network is robust to this switch, we conducted the same analyses looking for either a linear (i.e., first order) or non-linear (i.e., second order) effect of Edit Distance. Consistent with our Categorical analysis, there was no effect of Edit Distance on change in response in the language regions.

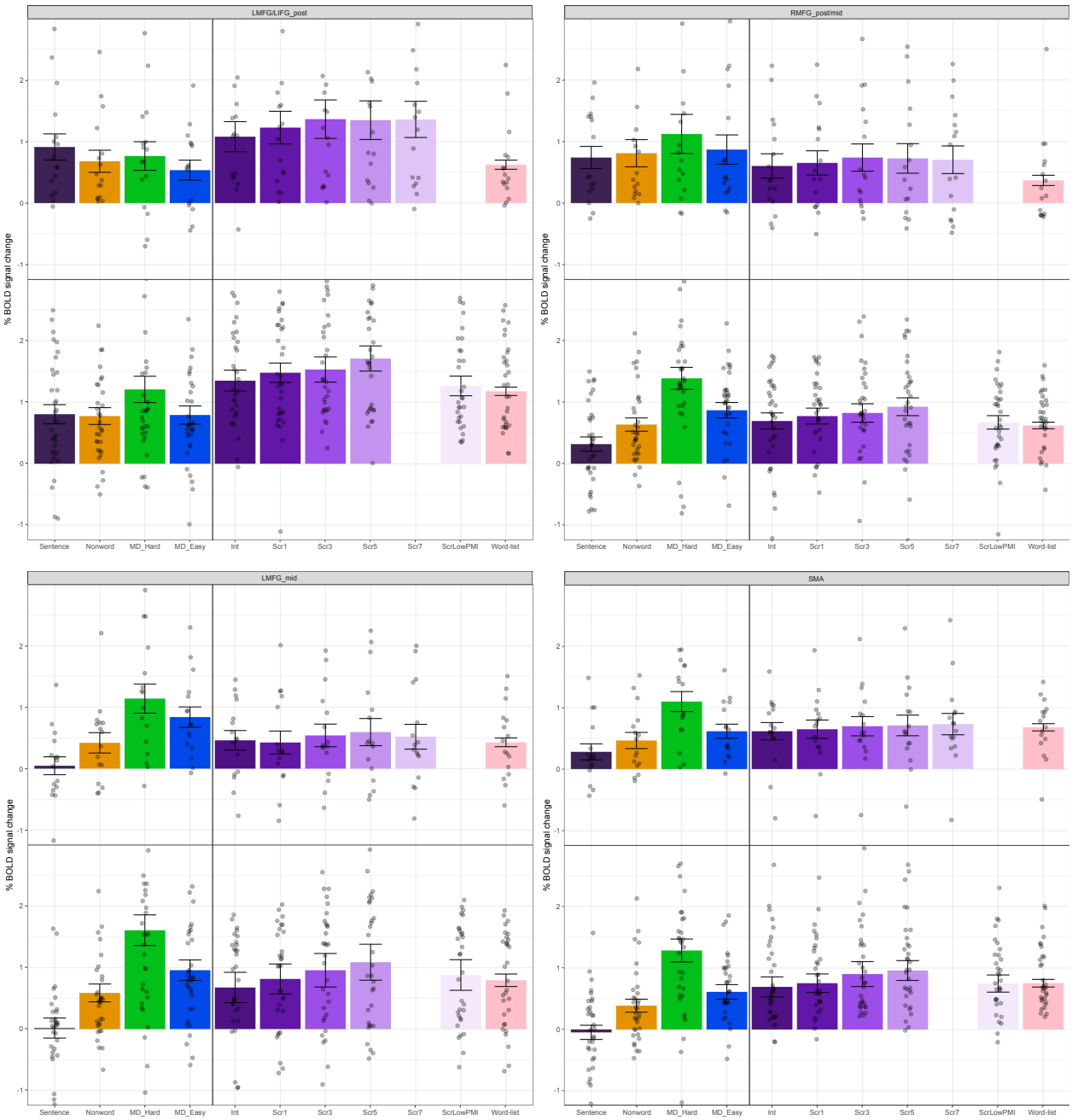

Figure S3: Neural responses (in % BOLD signal change relative to fixation) to the conditions of Experiments 1 (top) and 2 (bottom), as well as the language localizer and spatial WM task in each of the four scrambling-responsive fROIs.