The language network responds robustly to sentences across diverse tasks

*Ruimin Gao[1], Chandler Cheung[2], Matthew Siegelman[3], Alvincé L. A. Pongos[4,5], Hope H. Kean[6,7], Alyx Tanner[8], Evelina Fedorenko** [6,7] Anna A. Ivanova** [1]*

[1]School of Psychology, Georgia Institute of Technology
[2]Department of Computer and Information Science, University of Pennsylvania
[3]Department of Psychology, Columbia University
[4]University of California, Berkeley
[5]University of California, San Francisco
[6]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[7]McGovern Institute for Brain Research, Massachusetts Institute of Technology
[8]New York University

*** - co-senior authorship*

Abstract

A network of left frontal and temporal brain areas supports language comprehension and production, implementing computations related to word retrieval and combinatorial linguistic processing. Here, we ask: are computations over linguistic input driven in a bottom-up way, by the input, or in a top-down way, by task demands? Participants (n=52) read sentences and nonword lists under six task conditions, including passive reading, reading with a memory probe at the end, and tasks that require deep semantic engagement. The sentences>nonwords contrast isolated the same set of language-responsive voxels across all tasks; the locations of those voxels were participant-specific, highlighting the value of individual-specific functional localization. We therefore conclude that language localization is robust to task variation. We then examined responses to each task in these language-responsive voxels (the language network) and in the domain-general multiple demand (MD) network, known to respond to task demands. The language network responded robustly to sentences across all tasks, with somewhat higher responses to semantically engaging tasks. In contrast, the MD network responded to both sentences and nonwords in the presence of a demanding task, which warrants caution when examining brain responses to language tasks accompanied by task demands, as such tasks engage two independent systems. A multivariate analysis revealed that stimulus information is more easily decodable in the language network, whereas task information is more decodable in the MD network. These results suggest that the language and MD networks perform complementary functions during task-driven language comprehension, with the language network primarily extracting information from linguistic input and the MD network determining the appropriate response to the task.

1 Introduction

The human brain has a dedicated set of regions for processing language (Fedorenko et al., 2024). These regions, known as the language network, respond to a broad range of linguistic conditions: they are active during both production and comprehension, process written, spoken, and signed language, and respond to typologically diverse languages (Deniz et al., 2019; Fedorenko et al., 2010; Malik-Moraleda, Ayyash et al. 2022; Menenti et al., 2011; Neville et al., 1998; Regev et al., 2013; Scott et al., 2017; Silbert et al., 2014; Vagharchakian et al., 2012). However, the degree to which the language network is driven bottom-up by the linguistic inputs vs. top-down by task demands is not well understood.

Language production is almost always associated with a top-down demand—to express a particular meaning—driven either by internal goals or external demands (e.g., when answering a question). Language comprehension, on the other hand, may or may not be associated with top-down demands. Indeed, much evidence suggests that the language network is robustly engaged in passive, task-free comprehension paradigms (e.g., Bedny et al., 2011; Fedorenko et al., 2010; Lipkin et al., 2022; Malik-Moraleda, Ayyash et al. 2022; Scott et al., 2017). The latter observation is consistent with the fact that humans often extract meaning from language spontaneously, such as when overhearing a conversation or reading ads. Similarly, passive exposure to linguistic narratives elicits robust activity in the language network (Blank & Fedorenko, 2017; Fedorenko et al., 2010; Regev et al., 2013; Shain et al., 2022; Sueoka et al., 2024), similar to how people often listen to audio books and podcasts just for fun. But some evidence indicates that attention can affect responses in the language network, both positively and negatively. If a person is keenly interested in the topic of the text, their language areas will respond more strongly (Olson et al., 2024). On the flip side, when participants engage in a challenging perceptual task that diverts their attention from linguistic content, activity in the language network declines substantially (Cohen et al., 2021; Ivanova et al., 2021).

Understanding the extent to which task demands modulate responses in the language network is important both theoretically and practically. Theoretically, this knowledge can help understand where the language network falls in the cortical hierarchy. In particular, brain regions vary widely in the extent of their modulation by task context. Low-level sensory regions, such as the primary visual and auditory cortices, largely respond to the same stimulus in the same way regardless of task (Elhilali et al., 2007; Kay et al., 2008); their activity is only slightly modulated by attention (Jäncke et al., 1999; Poghosyan & Ioannides, 2008; Somers et al., 1999), and the auditory cortex responds to sounds even during sleep (Issa & Wang, 2008; Sela et al., 2016). Higher-level perceptual regions,

such as the inferotemporal visual cortex, are more sensitive to task demands (Harel et al., 2014; Koida & Komatsu, 2007; McKee et al., 2014; Sigala & Logothetis, 2002) but are still primarily driven by the inputs (McMains & Kastner, 2011, Bugatus et al., 2017, Keller et al., 2022) and are engaged robustly even when participants are focusing on unrelated tasks (Marvi et al., 2025). And some brain regions are strongly task-driven: for example, the fronto-parietal multiple demand (MD) network, implicated in goal-directed action and fluid reasoning, responds only during challenging cognitive tasks (Assem et al., 2020; Duncan & Owen, 2000; Zanto & Gazzaley, 2013). Does the language network resemble perceptual processing regions, which are driven primarily by the external inputs, or is it more similar to the MD network, which shows strong modulation by task demands?

On the practical side, a common way to identify the language network is with a passive comprehension contrast, such as simply reading sentences vs. nonword sequences (Fedorenko et al., 2010) or listening to sentences/short passages vs. to incomprehensible speech, such as unfamiliar-language speech, backwards speech, or acoustically degraded speech (e.g., Malik-Moraleda, Ayyash et al. 2022). The passive nature of the task makes it challenging to assess the level of stimulus engagement or the degree of comprehension other than indirectly, by the amount of activation in the language areas. However, including a task may engage brain regions that are not specifically engaged in language processing, but instead support general attention and working memory demands, such as the MD network introduced above. Indeed, a meta-analysis of 30 fMRI language paradigms by Diachek et al. (2020) revealed that the language network responds similarly strongly, on average, to passive comprehension conditions and to conditions with active tasks, but the latter paradigms additionally engage the MD network, which shows no response during passive comprehension. Given that the language and the MD networks have been established to be functionally distinct (Braga et al., 2020; Du et al., 2024; Fedorenko & Blank, 2020; Mineroff, Blank et al., 2018; Quillen et al., 2021), the use of language paradigms that include an active task may be undesirable, as such paradigms would conflate these two networks. However, all past comparisons of language paradigms with vs. without tasks have been indirect, performed across participants and across paradigms that differ in more than simply the presence or type of task, leaving some uncertainty as to how task demands affect brain responses to language.

Here, we systematically vary task demands in the same participants for the same materials across six tasks, while keeping other aspects of the paradigm as consistent as possible, in order i) to determine whether localizers that use different tasks produce comparable results; ii) to better understand the effects of tasks in language paradigms with respect to the engagement of the domain-general MD network, and more generally,

iii) to understand the language and the MD networks' contributions to language processing under different task conditions, and to clarify how the language network fits within the broader continuum of task-driven neural responsiveness.

To foreshadow our key findings, across all (six) tasks, the language network responded robustly to the sentences>nonwords contrast, but the sentence conditions with a task additionally engaged the MD network. The presence of a deeper, semantic task led to a stronger response to the sentence condition in the language regions, but based on the combination of univariate and multivariate analyses, we conclude that the response is primarily stimulus-driven, in sharp contrast to the MD network, whose activity scales with cognitive demand across stimulus types.

2 Methods

2.1 Participants

52 individuals (mean age = 23.9 years, SD = 3.8; 31 female) from MIT and the surrounding community participated for payment. 42 participants were right-handed and one was left-handed (but with typical, left-lateralized language responses), as assessed by Oldfield's (1971) handedness questionnaire; handedness information was not collected for the remaining 9 participants (see **Supplemental Figure 1** for the laterality index of the included participants). Six additional participants were tested, but excluded from the analyses: three due to excessive movement, two due to drowsiness, and one due to right-lateralized language responses. All participants had normal or corrected-to-normal vision. The study protocol was approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES), and all participants provided written informed consent in accordance with protocol requirements.

2.2 Design, materials, and procedure

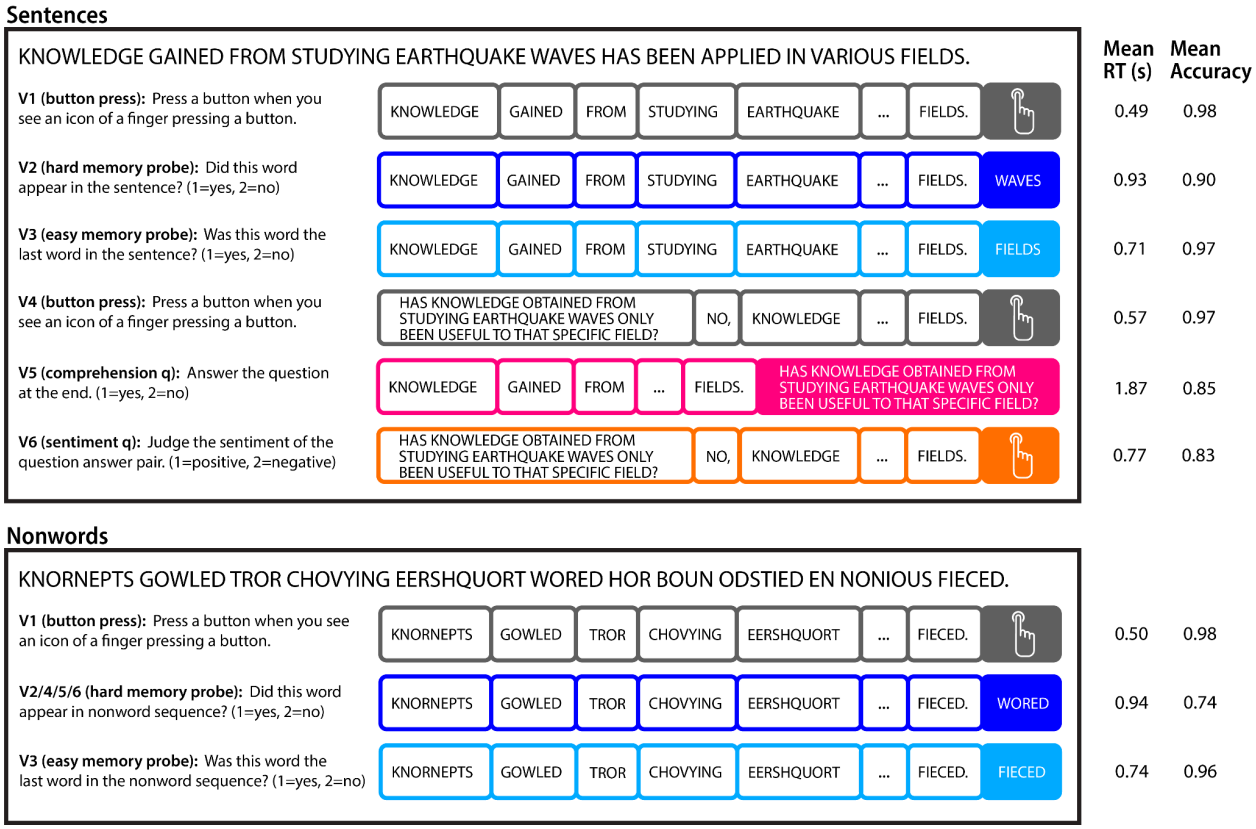

Figure 1. *Design of the experiment. The sentences and nonword strings were presented in the center of the screen, one word/nonword at a time, at the rate of 400 ms per word/nonword. In V4-V6, where a question was included, the question was presented all at once. In V1 and V4, participants were simply asked to press a button when they saw the finger icon appear at the end of the sentence. In V2 and V3, participants had to decide whether the probe word/nonword was present in the preceding sentence/nonword string; in V3, the correct probe was the same as the last word/nonword, which makes the task easy (effectively, deciding whether or not you just saw this word/nonword). For V1-V3, the task was identical between sentences and nonwords, for V4-V6, the hard memory probe version was used for the nonwords.*

Six different versions of the language localizer task were used in the current study (**Figure 1**). Each participant completed between 3 and 5 versions (with distinct sets of experimental materials; materials were rotated across task versions across participants), and each task version was completed by between 19 and 52 participants (**Table 1**).

Across all versions, the critical condition required participants to read 12-word-long sentences (e.g., KNOWLEDGE GAINED FROM STUDYING EARTHQUAKE WAVES HAS BEEN APPLIED IN VARIOUS FIELDS), and the control condition required them to

read lists of unconnected, pronounceable nonwords (e.g., KNORNEPTS GOWLED TROR CHOVYING EERSHQUORT WORED HOR BOUN ODSTIED EN NONIOUS FIECED). The sentences (a total of 5 sets, n=48 sentences each) were drawn from a language corpus (Francis & Kučera, 1964); the nonword lists were created by turning the words comprising the sentences into nonwords using the Wuggy software (Keuleers & Brysbaert, 2010). The materials are available at https://osf.io/zyfmu.

In three versions (V1-V3), the sentence trials consisted of a single sentence presented one word at a time (**Figure 1**).These versions have been commonly used in past work (e.g., Fedorenko et al., 2010; Hu, Small et al., 2023; Lipkin et al., 2022; Malik-Moraleda, Ayyash et al. 2022). In ***Version 1*** *(*V1, "button press"), each sentence/nonword list was followed by a screen showing a picture of a finger pressing a button and participants were asked to press a button whenever they saw this picture (for the timing details, see **Supplemental Table 1**). In Versions 2 and 3, each sentence/nonword list was followed by a memory probe word/nonword and participants had to press a button to decide whether the probe was present in the preceding stimulus (button 1 for yes, button 2 for no). In ***Version 2*** (V2, "hard memory probe"), the correct word/nonword probe could come from any position in the sentence/nonword list; in ***Version 3*** (V3, "easy memory probe"), the correct probe came from the last position, which effectively turns this task into a decision about whether the probe was the last word/nonword.

In the sentence condition in the remaining three versions, participants read question-answer pairs. ***Version 4*** (V4, "button press") was similar to V1, except that the sentence was preceded by a comprehension question presented on the screen all at once. The sentence was then presented (one word at a time, as in V1-V3, but with "Yes," or "No," appended to the beginning), such that the sentence provided the answer to the question. This version was designed as a control for Versions 5 and 6 that matches the amount of language material in the sentence condition. Versions 5 and 6 were designed to encourage deep engagement with the content of the sentences. In ***Version 5*** (V5, "comprehension q(uestion)"), each sentence was followed by a yes/no comprehension question (shown on the screen all at once, as in V4), and participants were asked to answer it (button 1 for yes, button 2 for no). In ***Version 6*** (V6, "sentiment q(uestion)"), the sentence trials were structured as in V4, but participants were instead asked to make a judgment about the emotional valence of the answer sentence (button 1 for positive content, button 2 for negative content). The nonword condition in V4-V6 used the hard memory probe task from V2.

The sentences and nonword lists were presented in the center of the screen, one word/nonword at a time, at the rate of 400 ms per word/nonword. Each trial was preceded by a 100 ms blank screen. A blocked design was used with 3

sentence/nonword-list trials per block. Each run consisted of 16 experimental blocks (8 per condition), and fixation blocks (14 s each, and included at the beginning of the run after every four experimental blocks). The order of sentence and nonword blocks was palindromic within the run and counterbalanced across runs. Across versions, run durations varied between 5 min 58 sec and 8 min 22 sec. For each version, a given participant completed two runs.

In addition to the critical language tasks, the majority of the participants (45 of the 52) completed a spatial working memory task, which was used to define the Multiple Demand system (Assem et al., 2020; Blank et al., 2014; Duncan, 2010; Mineroff, Blank et al., 2018). Briefly, participants viewed sequences of spatial locations flash up within a 3 × 4 grid, one at a time for a total of 4 locations (in the easy condition) or two at a time for a total of 8 locations (in the hard condition). At the end of each trial, they were shown two sets of locations and had to choose the set they just saw (see Supplementary Figure 3 for details).

Table 1. *Participant counts for each unique combination of completed language task versions.*

Task	V1	V2	V3	V4	V5	V6	# Participants
	X	X	X		X	X	18
	X	X	X		X		10
	X	X	X				1
	X	X		X	X	X	5
	X	X		X	X		1
	X	X		X		X	1
	X	X			X	X	1
	X	X			X		1
	X		X	X	X	X	6
	X		X	X	X		1
	X		X		X	X	1
	X			X	X	X	3
	X			X	X		1
	X			X		X	1
	X					X	1
Total per task	52	38	37	19	48	37	52

2.3 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 176 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 90o 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 the interleaved order (with 10% distance factor), 2.1 mm x 2.1 mm in-plane resolution, FoV in the phase encoding (A >> P) direction 200 mm and matrix size 96 mm x 96 mm, TR = 2,000 ms and TE = 30 ms. The first 10 s of each run were excluded to allow for steady state magnetization.

2.4 Behavioral performance in the scanner

We analyzed response times using a linear mixed-effects model and accuracy using a binomial mixed-effects model. Both models included fixed effects of stimulus type (sentence vs. nonwords), linguistic demand (question–answer pair vs. single sentence), task (button press, hard memory probe, easy memory probe, comprehension question, sentiment question), and the task × stimulus-type interaction. For both models, we included by-participant random intercepts and random slopes for task and stimulus type. For the response time model, we additionally included a random intercept for the material set, which significantly improved model fit ($p < .05$). For the accuracy model, including material set did not improve model fit ($p = 0.621$), so it was omitted.

To evaluate the accuracy of in-scanner sentence sentiment ratings, we collected responses from human participants on Amazon Mechanical Turk. We collected data from 100 participants; after filtering for accuracy on attention check questions, native language (English), and native country (United States), we retained data from 80 participants. Each participant rated 1/4 of the sentiment sentences. The resulting average was 19.96 responses per sentence.

Runs with accuracies that were below 63.1%, which is a value that falls three standard deviations below the overall mean (across participants, versions, and conditions) were excluded, except for cases where the buttons were swapped or responses were not recorded due to equipment malfunction (this led to the exclusion of 5 of the 472 runs). In addition, we ensured that each participant had two runs per version, so if one run within a version had below-threshold accuracy, we excluded the version altogether for that participant (this criterion led to the exclusion of 5 additional runs).

2.5 fMRI data preprocessing

fMRI data were analyzed using SPM12 (release 7487), CONN EvLab module (release 19b), and other custom MATLAB scripts. Each participant's functional and structural data were converted from DICOM to NIFTI format. All functional scans were coregistered and resampled using B-spline interpolation to the first scan of the first session (Friston et al., 1995). Potential outlier scans were identified from the resulting subject-motion estimates, as well as from BOLD signal indicators, using default thresholds in CONN preprocessing pipeline (5 standard deviations above the mean in global BOLD signal change, or framewise displacement values above 0.9 mm; Nieto 2020), and used as regressors of no interest in first-level analyses. Functional and structural data were independently normalized into a common space [the Montreal Neurological Institute (MNI) template; IXI549Space] using SPM12 unified segmentation and normalization procedure (Ashburner & Friston, 2005) with a reference functional image computed as the mean functional data after realignment across all timepoints omitting outlier scans. The output data were resampled to a common bounding box between MNI-space coordinates (−90, −126, −72) and (90, 90, 108), using 2-mm isotropic voxels and fourth-order spline interpolation for the functional data, and 1-mm isotropic voxels and trilinear interpolation for the structural data. Last, the functional data were smoothed spatially using spatial convolution with a 4-mm FWHM Gaussian kernel.

For all experiments, effects were estimated using a general linear model (GLM) in which each experimental condition was modeled with a boxcar function convolved with the canonical hemodynamic response function (HRF) (fixation was modeled implicitly, such that all timepoints that did not correspond to one of the conditions were assumed to correspond to a fixation period). Temporal autocorrelations in the BOLD signal time series were accounted for by a combination of high-pass filtering with a 128 s cutoff, and whitening using an AR (0.2) model (first-order autoregressive model linearized around the coefficient a = 0.2) to approximate the observed covariance of the functional data in the context of restricted maximum likelihood estimation. In addition to experimental condition effects, the GLM design included first-order temporal derivatives for each condition (included to model variability in the HRF delays), as well as nuisance regressors to control for the effect of slow linear drifts, subject-motion parameters, and potential outlier scans on the BOLD signal. For the univariate analyses, we modeled the data using condition-level regressors; for the multivariate decoding analyses, we extracted estimates for the individual blocks.

2.6 Defining language and MD functional regions of interest (fROIs)

Language and MD fROIs were defined using the group-constrained subject-specific (GSS) approach (Fedorenko et al., 2010; Julian et al., 2012) where a set of spatial masks, or parcels, is combined with each individual participant's localizer activation map, to constrain the definition of individual fROIs. The parcels delineate the expected gross locations of activations for a given contrast based on prior work and large numbers of participants, and are sufficiently large to encompass the inter-individual variability in the locations of functional areas. For the language fROIs, we used a set of ten parcels derived from a group-level probabilistic activation overlap map for the sentences > nonwords contrast in 220 participants and used in much prior work (e.g., Malik-Moraleda, Ayyash et al. 2022; Olson et al., 2024; Schrimpf et al., 2021; Shain et al., 2022; Tuckute et al., 2024). For each hemisphere, the parcels included two regions in the inferior frontal gyrus (IFG, IFGorb), one in the middle frontal gyrus (MFG), and two in the temporal lobe (AntTemp and PostTemp). For the MD fROIs, we used a set of 20 parcels (10 in each hemisphere) derived from a group-level probabilistic activation overlap map for the hard > easy spatial working memory contrast in 197 participants and also used in many prior studies (e.g., Assem et al., 2020; Kean et al., 2025; Shain et al., 2022). The parcels included regions in the frontal and parietal lobes, as well as a region in the anterior cingulate cortex. The language and the MD parcels are available at https://www.evlab.mit.edu/resources-all/download-parcels. Within each parcel, we selected the 10% of voxels with the highest t-values for the contrast of interest (sentences > nonwords for the critical language tasks, hard > easy spatial working memory for the MD localizer).

2.7 Response magnitude analyses

We evaluated the responses of the language and MD fROIs to the conditions of the critical language tasks. For each condition, the effect was averaged across the voxels within each fROI to derive a single value per fROI per participant. When estimating the responses to the conditions of the task that was used as the localizer for that fROI, the effects were estimated using across-runs cross-validation to ensure independence (Kriegeskorte et al., 2009).

2.8 Spatial overlap analyses

To assess the degree of overlap between language fROIs defined using different localizer task versions, we used the Dice coefficient measure (Dice, 1945). To equalize the amount of data between the within-task comparisons (where each data half consists of a single run) vs. across-task comparisons (where each task consists of two runs), for

the latter, we evaluated the overlap between fROIs defined by localizer task A (run 1) and localizer task B (run 2), and vice versa, averaging across the splits to derive a single value per fROI per participant. In the analyses, we compared the within-participant between-task consistency to the within-task between-participant consistency.

2.9 Spatial correlation analyses

To investigate the activation patterns of voxel-level responses within each language fROI, we extracted vectors with that fROI's voxelwise responses to a single stimulus type (sentences/nonwords) and a single language task version (2 stimulus types x 6 task versions = 12 conditions in total). We then computed Pearson correlation values for all pairwise condition combinations for each fROI. To ensure the independence of data used to estimate the voxelwise responses in each condition, we used separate runs to calculate each of the two effects we compare (and, when possible, averaging between estimates for (a) Run 1 - Condition 1 vs. Run 2 - Condition 2 and (b) Run 1 - Condition 2 vs. Run 2 - Condition 1). Because we have a total of 2 runs per task version, when we use one run as the fROI localizer, we do not have enough remaining data to compare two effects from the same task version. As a result, we do not provide values for V1 Sentence/Nonword - V1 Sentence/Nonword correlation when using V1 as the localizer. The resulting Pearson correlation values were Fisher z-transformed for regression analyses.

2.10 Multivariate decoding of stimulus vs. task information

To investigate whether different stimulus types and different tasks evoke distinct, linearly separable response patterns in the language and the MD networks, we trained within-participant linear classifiers on block-level voxel responses to each condition. For this analysis, we only included V1, V2, and V3, because the task was identical between the sentence and nonwords conditions for these versions. To ensure that the results were not biased by the localizer version used to define the fROIs, we used V4 as the localizer. A total of 29 participants completed all V1, V2, and V3 tasks; of these, 21 also completed V4 and the MD localizer. These 21 participants were included in this analysis.

We normalized the data across the voxels within each fROI, for each block separately, to remove univariate effects that distinguish conditions. We adopted five-fold cross-validation across blocks to split the data into train and test sets. To ensure generalizability, we evaluated the performance of four classifiers—a correlation-based classifier, k-nearest neighbors (kNN), logistic regression, and support vector machines (SVM)—and the results were similar across them. We report the results from the

correlation-based classifier in the main text (see **Supplemental Table 12** for the results of the other approaches).

2.11 Statistical analyses

We conducted the analyses using mixed-effects models implemented in the R package lmerTest (Kuznetsova et al., 2017). Separate models were fit for the left and right hemisphere language and MD systems.

Response magnitude analyses: We fit a mixed-effects model predicting fMRI activation (BOLD signal change) with fixed effects of linguistic demand (which varied across task versions (single sentence (used in V1-V3, and question-answer pair (used in V4-V6)), the interaction between the localizer task and linguistic demand to test whether responses to the sentence condition are modulated by task demand, and a three-way interaction between the localizer task, linguistic demand, and stimulus type (sentences, nonwords) to test whether nonword responses differ significantly from sentence responses for each localizer task. We included random intercepts for fROIs and participants, and random slopes for condition (sentences, nonwords) and task version where possible. For language-network analyses, we additionally included a random intercept for localizer-task version. When comparing the results across localizer versions used to define the fROIs, we fit separate models for each localizer version; and when comparing the results across fROIs, we fit separate models for each fROI; in both cases, an FDR correction was used to account for multiple comparisons.

Spatial overlap analyses: We fit a mixed-effects model predicting spatial overlap with fixed effects of comparison type (within-participant vs. between-participant) and task (within-task vs. between-task), and their interaction. Random intercepts were included for participant pairs and fROIs. Post-hoc comparisons with Tukey's correction were conducted to examine: (1) within-participant within-task vs. within-participant between-task overlap for each task version, and (2) within-participant between-task vs. between-participant within-task overlap for each task version.

Spatial correlation analyses: We fit a mixed-effects model predicting Fisher-z-transformed correlation values with fixed effects of effect pair (sentences vs. sentences, sentences vs. nonwords, nonwords vs. nonwords), the interaction between linguistic demand (single sentence, question-answer pair), and the interaction between effect pair and task presence (no task (V1, V4), task (V2-V3, and V5-V6)). Random intercepts were included for participants and fROIs in all models; for language-network analyses, we additionally included a random intercept for localizer-task version.

Decoding analyses: To test the significance of decoding accuracy, we conducted permutation tests (N = 1,000) to determine whether participants' average decoding accuracies exceeded chance with an FDR correction. To compare decoding accuracy across networks, we fit a binomial mixed-effects model for prediction accuracy that included a fixed three-way interaction among training type (within- vs. across-task for stimulus type decoding, or within vs. across-stimulus-type for task decoding), network (LH language network vs. MD network), and testing group (task version (V1/V2/V3) for stimulus-type decoding or condition (sentences/nonwords) for task decoding). Random intercepts were included for participants and classifier type (correlation-based/kNN/logistic/SVM). Post-hoc comparisons with a Bonferroni correction were conducted to examine accuracy differences across networks for each combination of training type and testing group.

3 Results

3.1 All six task versions of the language localizer are effective at identifying the language areas within individual brains.

First, we examined the responses to sentences and nonwords in the language network defined using different task versions of the localizer, and tested whether these responses were spatially consistent across fROIs (**Figure 2**). For all language task versions, the left-hemisphere language network showed a significantly stronger response to sentences compared to nonwords (ps < .001), regardless of the localizer used for fROI definition (see **Supplemental Table 2** for the statistical results). These effects also held for all five left-hemisphere language fROIs individually (ps < .001; **Supplemental Table 3**). These results replicate prior work demonstrating robust sentence > nonword contrasts in the language network (See Fedorenko et al., 2024 for review), and extend it by showing that these effects hold across six different task versions.

Next, we assessed the topographic consistency of the language fROIs defined using the same or different tasks within individuals, or—for comparison—using the same task across individuals. Within-participant overlap values were consistently high, both within each task (0.59 on average; **Figure 3**), consistent with prior reports of topographic stability in the language network (e.g., Mahowald & Fedorenko, 2016), and importantly across tasks (0.50 on average), extending prior findings of similar activations across different versions of the language localizer varying in modality and materials (e.g., Malik-Moraleda, Ayyash et al. 2022; Olson et al., 2024; Scott et al., 2017). These values are considered to fall in the moderate-high range (Wilson et al., 2017). The within-task

patterns did show significantly greater similarity compared to the between-task patterns, which suggests that different task variants activate slightly different sets of voxels (ps < .001, **Supplemental Table 10**). However, even the within-individual between-task values were substantially higher than the between-individual within-task values (0.13 on average, which is considered to be in the low range; Wilson et al., 2017) (ps < .001), consistent with earlier findings of substantial inter-individual variability in the precise locations of the language areas (Braga et al., 2020; Du et al., 2024; Fedorenko et al., 2010; Mahowald & Fedorenko, 2016; Shain & Fedorenko, 2025). These patterns were consistent across all five left-hemisphere language fROIs (**Supplemental Table 11**).

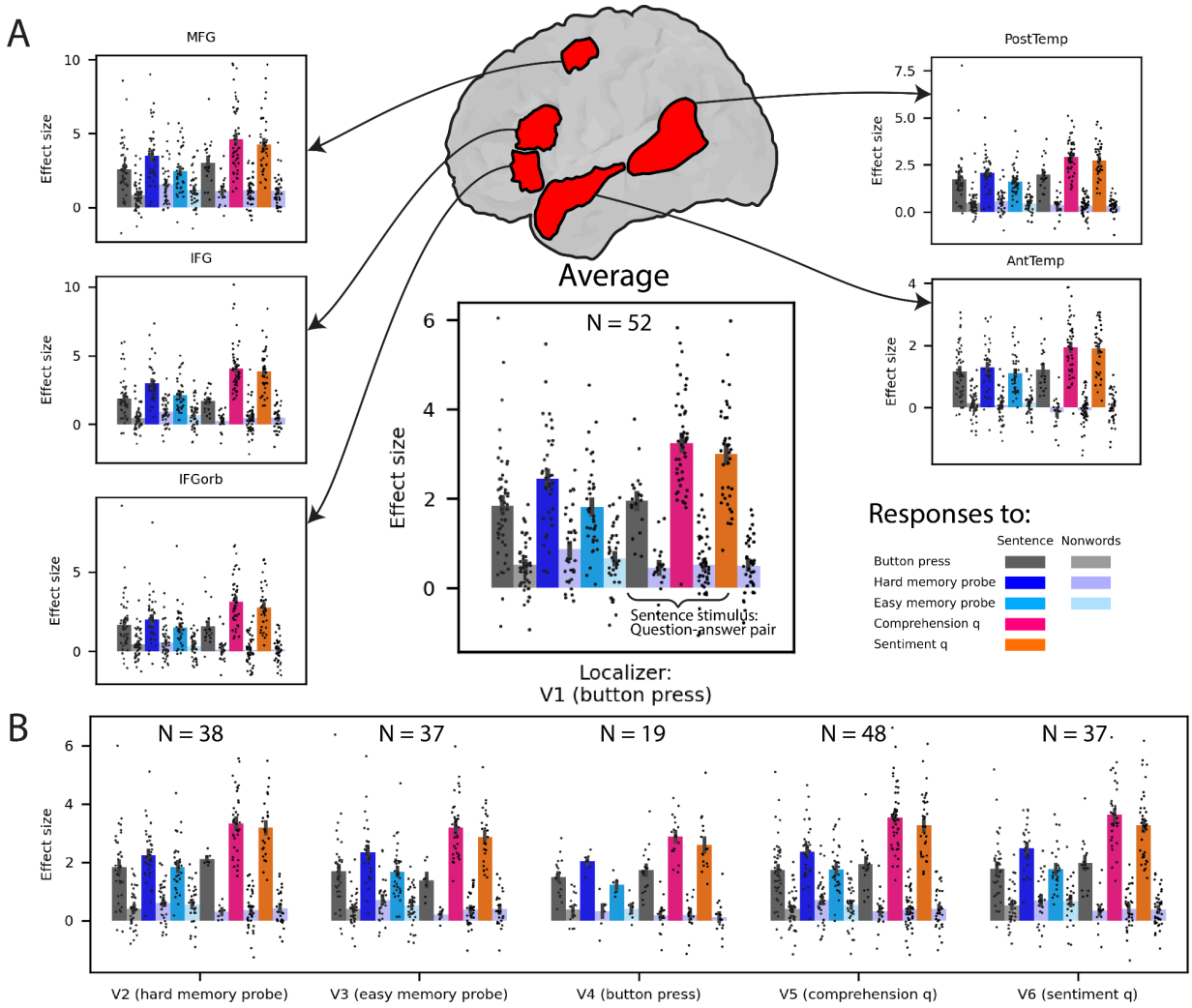

Figure 2. *Responses of the core left-hemisphere language network to the sentence and nonword conditions in the six task versions. (A) Average network-level response profile (large bar graph) and fROI-level profiles (smaller bar graphs). The language fROIs were defined with language localizer V1 (button press), which every participant performed. Error bars denote the*

standard errors of the mean by participant; dots correspond to individual participants. (B) Average network-level response profiles for language fROIs defined using the other five localizer versions (V2-V6).

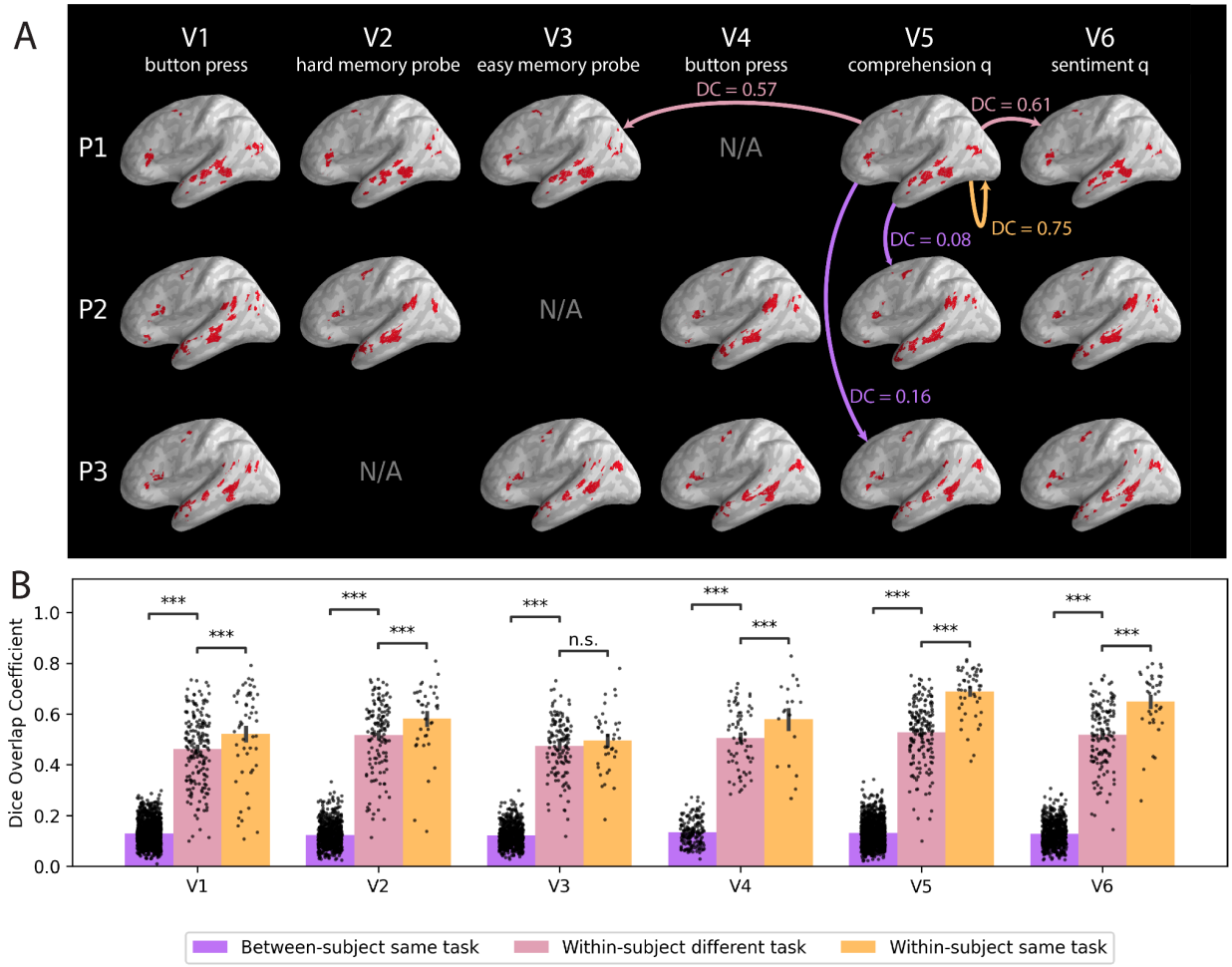

Figure 3. *Language fROIs defined by different language localizer task versions. (A) Language fROIs defined in three sample participants (P1-P3). Each participant completed 5 of the 6 localizer versions. DC = Dice overlap coefficient. The Dice coefficient is calculated for the left-hemisphere language fROIs within participants within and between tasks (orange and pink arrows) or between participants for the same task (purple arrows) (Methods). (B) Average Dice overlap of left-hemisphere language fROIs within participants (orange and pink bars) and between participants for the same task (purple bar) (ns: not significant, *: p < .05, **: p < .01, ***: p < .001). Separate runs are used to define each of the two fROIs being compared (averaging between estimates for (a) Task 1 - Run 1 vs. Task 2 - Run 2 and (b) Task 1 - Run 2 vs. Task 2 - Run 1). Error bars denote the standard errors of the mean over participants / participant pairs; dots correspond to participants (for the within-participant comparisons) and to pairs of participants (for the between-participant comparison).*

3.2 Task demands lead to higher responses in the language network but also engage the multiple demand network.

The language and the multiple demand network are differentially affected by stimulus type vs. task in both univariate responses and activation patterns.

We investigated the effects of stimulus type and task on responses in the left and right hemisphere components of the language and MD networks. The two networks showed distinct activation profiles (**Figure 4A**), consistent with prior findings (Blank et al., 2014; Fedorenko et al., 2013; Hiersche et al., 2024; Mineroff, Blank et al., 2018). In the language network, we observed a robust sentence > nonword contrast for all tasks in both hemispheres (ps < .001; **Supplemental Table 4**). In contrast, the MD network showed higher responses to nonwords than to sentences in both hemispheres for V1-V4 (ps < .05); for V5 (comprehension question) and V6 (sentiment question), the sentence and nonwords conditions did not reliably differ. An important thing to note is that all sentence conditions with an active task (i.e., tasks requiring more than a simple button press at the end of a stimulus) elicited an above-baseline response in the MD network (ps<.05), and in some cases, this response was higher than the response to the nonwords condition. This means that depending on the control condition, sentence comprehension with an active task may engage not only the language areas but also the MD network.

Regarding task effects, the left-hemisphere (LH) language network showed increased activation for the sentence condition in V2 (hard memory probe) relative to the button-press task (V1) (p < .001). No difference was observed for this contrast in the RH language network, nor for the comparison of V3 (easy memory probe) relative to V1 for either the LH or the RH language network. Further, both the LH and RH language network showed increased activation for the sentence condition in V5 (comprehension question) and V6 (sentiment question) relative to the button-press version that was matched for the amount of linguistic input (V4), as well as relative to the sentence conditions in V1-V3 (all ps < .001). Importantly, however, these task-related effects were dwarfed by the effects of stimulus type: a 5.08 times higher response to the sentence condition compared to the nonword-list in the LH language network (5.92 in the RH language network).

In contrast, the MD network exhibited more robust task effects compared to the stimulus-type effects: in both hemispheres, for all “active” tasks, the sentence condition elicited a higher response than the corresponding button-press condition matched for the amount of linguistic content (i.e., V2 and V3 compared toV1; and V5 and V6 compared to V4) (all ps < .001). Furthermore, across the 12 conditions, the average response in the MD network was well predicted by the difficulty of that condition as

reflected in average accuracies (ps < .05 for both LH and RH) and reaction times (p < .05 in the LH; **Supplemental Figure 2**). Such correlations were small and not significant for the language network's responses (all ps > .30).

The univariate responses were mirrored by the fine-grained multivariate activation patterns. Within both LH and RH language networks, the activation pattern was more consistent among the sentence conditions than among the nonword conditions (ps < .001; **Figure 4B**; **Supplemental Table 7**). However, for the MD network, this effect was observed only in the left hemisphere (p < .001; **Supplemental Table 8**). Furthermore, in both LH and RH language networks, the activation pattern for the sentence condition became more distinct from that for the nonword condition when the sentence conditions involved the additional linguistic demand of a question-answer pair (in V4-V6) (ps < .001). In contrast, in the MD network, the increased linguistic demand in the sentence conditions resulted in activation patterns that were more similar to those for the nonword conditions (ps < .001). Finally, both networks showed more consistent within-condition activation patterns under task conditions compared to passive reading (button press) versions matched for the amount of linguistic content (i.e., V2 and V3 vs.V1; and V5 and V6 vs. V4) (ps < .001). However, in the LH language network, this effect was 2.80 times smaller than the sentences > nonwords effect (1.10 smaller for the RH language network), but in the LH MD network, it was 2.51 times greater than the sentences > nonwords effect (15.6 greater for the RH MD network).

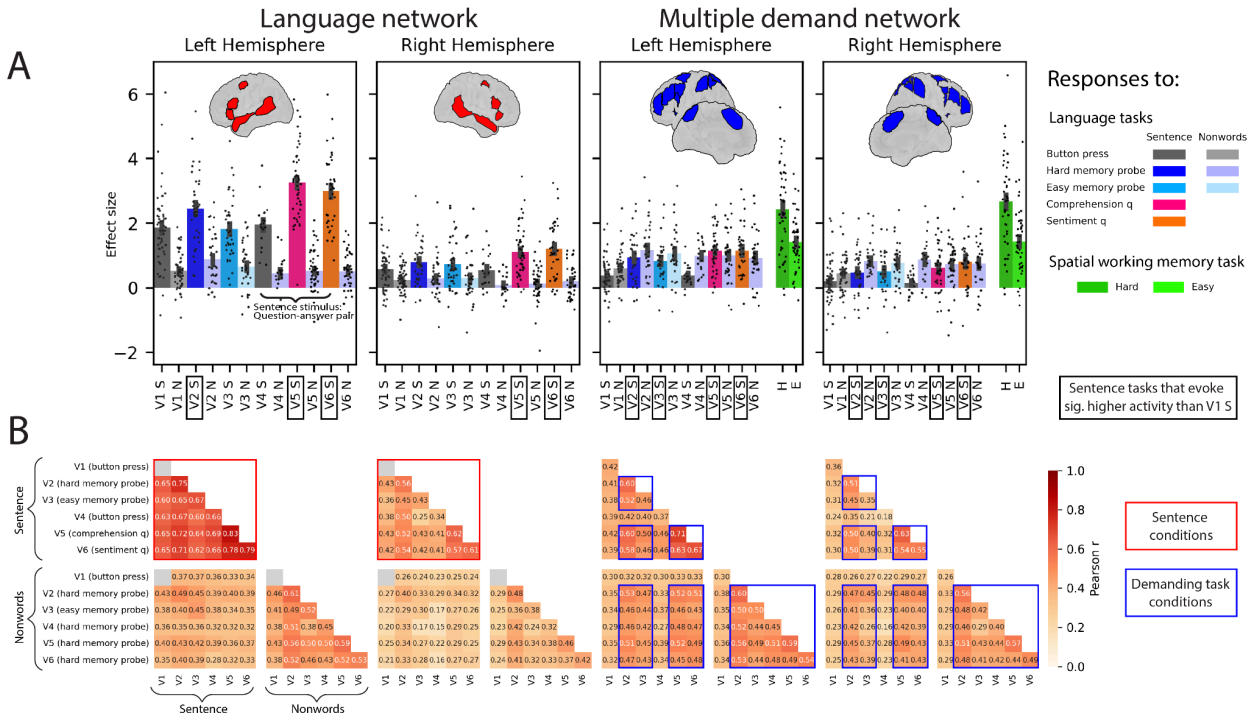

Figure 4. *(A) Responses of the left-hemisphere and right-hemisphere language and MD networks to the sentence and nonword conditions in the six task versions. As in Figure 2, here and in the results in panel The language fROIs were defined with language localizer V1 (button press), which every participant performed. Error bars denote the standard errors of the mean by participant; dots correspond to individual participants. (B) Spatial correlation analyses of activation patterns within the networks. Separate runs are used to estimate each of the two effects being compared (averaging between the estimates for (a) Task 1 - Condition 1 vs. Task 2 - Condition 2 and (b) Task 1 - Condition 2 vs. Task 2 - Condition 1) .*

3.3 Fine-grained activity in the language network better distinguishes stimulus type but activity in the Multiple Demand network better distinguishes task type.

Finally, we examined how well stimulus type (sentences vs. nonwords) and task (button press vs. hard memory probe vs. easy memory probe) could be decoded from the activation patterns in the language and MD networks. Both stimulus type and task could be reliably decoded in the LH language network and in the MD network, whether the classifier was trained and tested on the same task or trained on two tasks and tested on the left-out task (ps < .05; **Figure 5**; **Supplemental Table 12**). However, post-hoc tests revealed that, for both within- and between-task classification, the language network

exhibited higher decoding accuracy for stimulus type compared to the MD network (average decoding accuracies 0.69 and 0.56, respectively; $p < .001$). In contrast, the MD network showed higher decoding accuracy for tasks than the language network, although the difference was less pronounced (0.42 vs. 0.40; $p < .001$, except when both training and testing were performed on the sentence condition: $p = .072$). Thus, although each network contains information about both stimulus type and task, they differ in how robustly they represent these two kinds of information, with stronger stimulus coding in the language network and stronger task coding in the MD network.

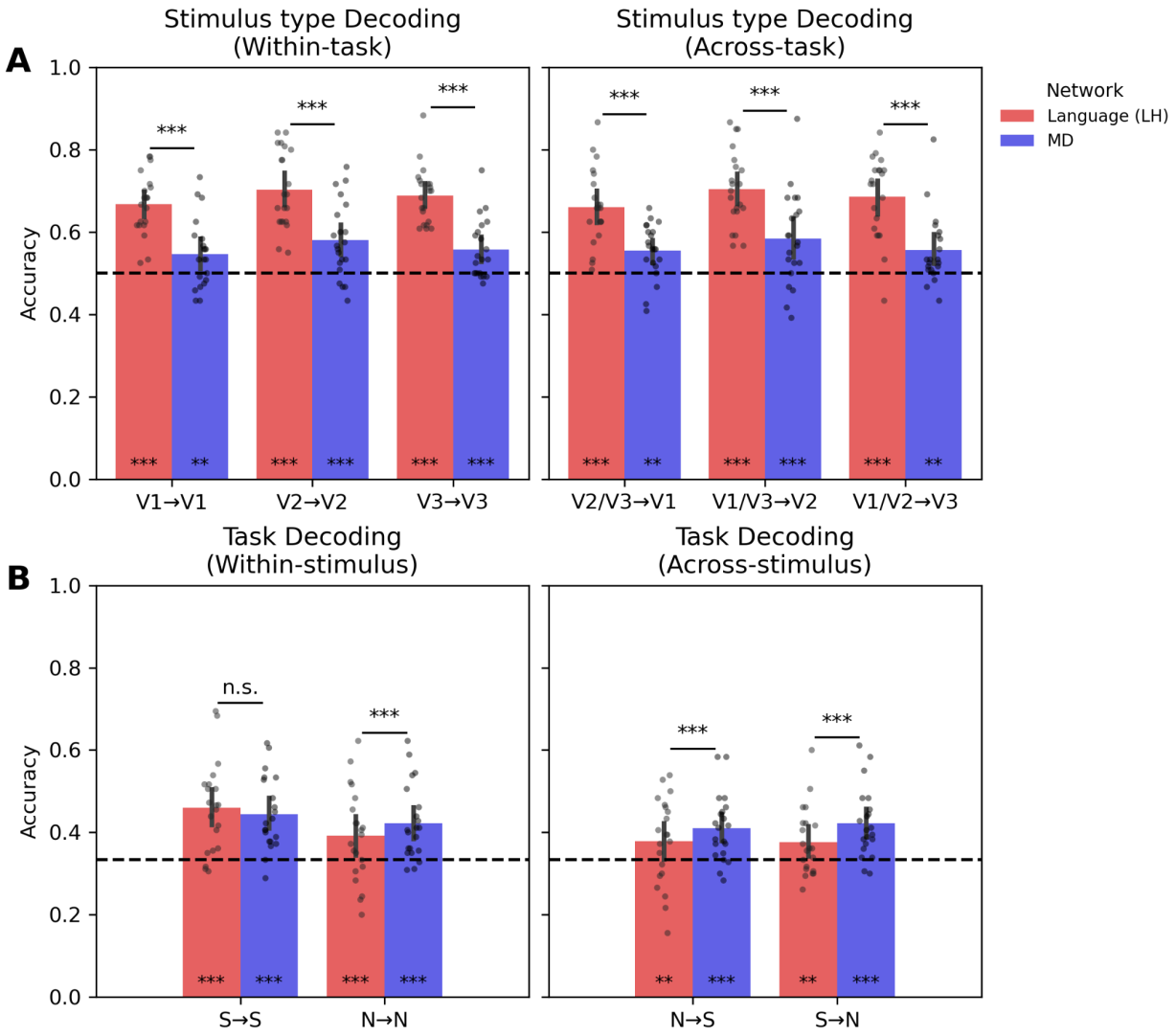

Figure 5. *Multivariate pattern analysis results for the LH language network (red bars; see Supplemental Table 12 for the results of the RH language network) and the bilateral MD network (blue bars) in classifying stimulus type (A) and task (B). (See Supplemental Table 12 for evidence that the results generalize across other types of classifiers.)*

4 Discussion

In this study, we systematically varied task demands—from a simple button-press to challenging memory probes and content-based comprehension questions/judgments—to test whether different task demands affect the localization and functional response profile of the language network in individual brains. We found that the language network localizer is robust to task variation, yielding a spatially consistent set of brain areas with remarkably similar activation profiles. Tasks with stronger attentional demands elicit somewhat stronger responses in the language areas, but they also engage the task-driven multiple demand (MD) network. Finally, the effects of stimulus type are stronger in the language than in the MD network, as reflected in both univariate responses and multivariate activation patterns, and the effects of tasks are stronger in the MD network. Below, we discuss the theoretical and practical implications of these findings.

4.1 The language network is largely driven bottom up, by stimulus properties

We found that the language network's response to sentences remained robust across task contexts. Specifically, the language areas showed a consistently stronger response to sentences than to nonwords regardless of whether participants were simply engaged in attentive passive reading (in our button-press conditions), in tasks that required them to remember the words that composed the sentences (in the memory-probe conditions), and in tasks that required evaluating the semantic content of the sentence (in the comprehension-question and sentiment-judgment conditions). Across the network and in each individual language region, the sentence condition elicited a response that was 4-8.5 times higher than the nonwords conditions. Moreover, the spatial locations of these language-selective areas were highly similar across task versions within individuals (and much more similar than for the same task version across individuals, in line with previously reported substantial inter-individual topographic variability; Braga et al., 2020; Fedorenko et al., 2010; Shain & Fedorenko, 2025). These results a) suggest that the response in the language network predominantly depends on the presence of particular linguistic features, and b) align with prior evidence of sensitivity of the language areas to cognitive demands related to word recognition and combinatorial syntactic and semantic processing (e.g., Blank et al., 2016; Fedorenko et al., 2010; Giglio et al., 2022; Pallier et al., 2011; Shain, Kean et al., 2024).

Although the language network's sensitivity to linguistic input was the dominant factor, we also detected moderate increases in activation during the more cognitively demanding tasks, such as those requiring memory retrieval or interpretive judgments. These task-induced modulations suggest that top-down influences can enhance the language system's engagement, but they do not override its fundamentally stimulus-driven response profile. Furthermore, it is difficult to rule out the possibility that task demands lead to the re-processing of the stimulus (e.g., to remember whether a word appeared in the sentence, one may reproduce the sentence in their mind), so the increased responses in the language network may be due to this additional linguistic processing rather than the performance of the task itself.

4.2 Implications for functional localization of the language network

The significant variability in the precise locations of the language regions across individuals poses a problem for approaches that average data voxelwise across individuals (see Fedorenko, 2021 for discussion). One alternative is to define these regions functionally in each individual brain using a simple 'localizer' task (Saxe et al., 2006)—this approach has been standardly used in vision research for many years (e.g., Downing et al., 2001; Epstein & Kanwisher, 1998; Kanwisher et al., 1997) and has been extended to many other domains, including language (Fedorenko et al., 2010). Most common approaches use a contrast between language processing and a perceptually matched control condition (e.g., reading sentences vs. lists of nonwords; or listening to sentences/passages vs. to backwards speech), typically accompanied by a button press or a memory probe task to ensure engagement (Bedny et al., 2011; Fedorenko et al., 2010; Hiersche et al., 2024; Lipkin et al., 2022; Malik-Moraleda, Ayyash et al. 2022; Scott et al., 2017). Here, we investigated whether or how task demands affect the locations and functional profiles of the language regions.

Our findings show that within a given individual, the topography of the functional regions of interest (fROIs) defined by the reading-based sentences>nonwords contrast is highly consistent across six task versions—a testament to the robustness and generalizability of this localizer contrast. In line with this topographic similarity, the response profiles of these fROIs are similar, all showing a large and replicable sentences>nonwords effect in a left-out run and selectivity relative to a demanding non-linguistic task (the spatial working memory task). These results suggest that researchers can use any of these versions of the localizer without compromising the validity or reproducibility of language network localization.

This robustness to task adds to prior evidence of the robustness of this localizer contrast (between language processing and a perceptually similar control condition

where the content is indecipherable) to modality of presentation (reading, listening, or audio-visual; Fedorenko et al., 2010; Olson et al., 2024; Scott et al., 2017), speed of presentation (Tuckute et al., 2024), language (showing generalizability across 45 spoken languages: Malik-Moraleda, Ayyash et al. 2022; Wolna et al., 2024; and sign languages; Richardson et al., 2020; Terhune-Cotter, 2025), materials (Casto et al., 2025; Scott et al., 2017; see also Supplemental Figure 4) and various experimental procedures (Lipkin et al., 2022). Importantly, this study reinforces prior findings that a passive language comprehension version of the localizer works well for identifying the language areas (e.g., Fedorenko et al., 2010; Olson et al., 2024; Scott et al., 2017). Of course, in some cases, having a behavioral task—to ensure deep engagement for participants who may struggle with attention, or to directly measure how well the stimuli are understood—may be desirable. In such cases, or when creating new versions of a language localizer, it is critical to ensure that the contrast is not difficulty-confounded. For example, some past studies have used a contrast between listening to sentences and answering questions about the content vs. listening to sentences played backwards and deciding whether a beep is present at the end (e.g., Newport et al., 2022; Olulade et al., 2020). Such difficulty-confounded contrasts will engage both the language network, but additionally the domain-general Multiple Demand network, which is sensitive to effort across diverse tasks (MacGregor et al., 2022). Given that these two networks support distinct cognitive functions, and the MD network plays a limited role in language processing (Shain et al., 2022), conflating them is undesirable and will lead to interpretive challenges.

Finally, it is worth noting that the language network can be identified within individual brains without any functional contrast, from patterns of functional connectivity alone. Braga et al. (2020; see also Du et al., 2024) have shown that this approach works for resting state data; and Du et al. (2025) and Shain & Fedorenko (2025) have extended this result to functional connectivity patterns extracted from random task data. So, those who remain skeptical about functional localizers may prefer to use these task-free data-driven ways of identifying the language areas. However, using localizers remains the most time-efficient approach: a single ~5 minute run (or even less; Lee et al., 2024) is sufficient in almost all cases.

4.3 The MD network is robustly modulated by task demands

Across analyses, the MD network showed the characteristic signature of a domain-general system sensitive to general cognitive effort. First, in the ***univariate analyses***, responses increased with task difficulty across different linguistic content (sentences vs. nonwords). Harder tasks, as measured behaviorally (i.e., hard and easy memory probe, comprehension question, sentiment judgment), elicited substantially

stronger activation in the MD network than the easier (button-press) tasks. Further, behavioral difficulty (accuracies, response times) reliably predicted MD responses across conditions. These results add to the growing body of evidence that the MD network is robustly engaged during diverse demanding tasks, including classic executive function tasks (Assem et al., 2020; Duncan & Owen, 2000; Fedorenko et al., 2013; Shashidhara et al., 2019) but also tasks in various domains of perception and cognition that may be associated with executive demands. For example, when language processing is accompanied by task demands above and *beyond linguistic processing* demands related to word recognition or syntactic parsing, the MD network shows above-baseline activity (e.g., Diachek et al., 2020; Quillen et al., 2021). Importantly, linguistic demands are supported by the specialized language network (Quillen et al., 2021; Shain et al., 2022; Wehbe et al., 2021).

And in the ***multivariate analyses***, the within-task activation patterns were more reliable for more demanding tasks, cross-task pattern similarity reflected task difficulty rather than stimulus type, and task was more robustly decodable from the MD network's activation patterns than from those in the language network. Although the activations in the MD network allowed for weak decoding of stimulus type, the activations were dominated by task information. These findings align with prior work showing that the MD network can adaptively code both task and stimulus information (Woolgar et al., 2011), and with findings from the visual domain where task information was also shown to be more strongly present in the MD network, whereas stimulus information was more strongly represented in high-level visual regions (Shashidhara et al., 2024). We extend this pattern to language: linguistic stimulus information is primarily discriminated in the language network, but the task performed on those stimuli is strongly represented in the MD network.

Taken together, the results reinforce the view that the MD network implements task-general computations that operate independently of the language network. When language processing is embedded within a cognitively demanding paradigm, MD regions become engaged—alongside the language network— to represent task structure and support task-relevant demands, while the language network performs specifically linguistic computations. This division of computational labor explains why language paradigms that include additional task demands systematically recruit two distinct networks and thereby risk conflating stimulus-driven language responses with demand-driven executive responses.

4.4 Limitations

Several limitations of the current study suggest clear directions for future work. First,our task battery—although the largest in evaluating effects of task on language network

localization—captures only a small fraction of the behaviors that constitute functional linguistic competence (Mahowald, Ivanova et al., 2024), i.e. the use of language in real-life contexts. All six tasks in our study used isolated sentences or question-answer pairs. The memory-probe tasks focused on the form of the sentence (asking about particular words comprising it), and the two higher-level tasks (comprehension questions, and sentiment judgments) required deeper engagement with the sentence's meaning but did not require discourse-level integration, reference resolution, pragmatic inference, domain-specific or abstract reasoning, or planning, which are central components of functional language use. Evaluating language processing in more complex paradigms would tell us whether the stable stimulus-driven signature of the language network persists when linguistic processing is embedded in broader reasoning or communicative contexts.

Second, our sample consists of young, neurotypical adults. As a result, extending these findings to more diverse populations—including children and older adults, multilingual speakers, and clinical populations—remains an important future goal.

Finally, fully dissociating stimulus and task processing is challenging: tasks that require additional memory or inference plausibly encourage participants to reprocess or rehearse the sentences, which may lead to additional linguistic computation. As a result, an increase in the response of the language network for more demanding tasks (as we observed) is difficult to interpret unambiguously. Designing experiments that minimize this ambiguity will require some creativity.

5 Conclusion

This study demonstrates that the language network exhibits robust, stimulus-driven responses to sentences regardless of task context. Across six task variants—including passive reading, memory probes, comprehension questions, and sentiment judgments—language-selective regions responded consistently more strongly to sentences than to nonwords. Furthermore, the language fROIs defined using any of these tasks were topographically almost as similar as when using different portions of the data from the same task, supporting the reliability and generalizability of functional localization methods for mapping the language network. Although task demands modulated the magnitude of responses within the language network, the activation pattern remained largely stable, suggesting that these regions are consistently engaged in linguistic input processing—demands that were stable across tasks.

These findings clarify the nature of task modulation in the language network: top-down influences, such as attention or memory demands, can amplify responses but do not fundamentally shift the topography or representational structure of the language-selective regions. Instead, a distinct network—the domain-general MD network—is preferentially engaged by cognitive task demands and strongly encodes task identity, while the language network primarily encodes linguistic content. In this way, the language network appears akin to perceptual brain areas, which are primarily input-driven.

Data and Code Availability

The experiment scripts, data, codes are available at https://github.com/RuiminGao/DiffTasks.git. The language and the MD parcels are available at https://www.evlab.mit.edu/resources-all/download-parcels.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

We acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research, MIT, including the technical support team (Steve Shannon and Atsushi Takahashi). We thank EvLab members for assistance with data collection and LIT lab members for feedback on this work. RG and AI were supported by Georgia Tech School of Psychology startup funds. EF was supported by research funds from the McGovern Institute for Brain Research, MIT's Simons Center for the Social Brain, MIT's Poitras Center for Psychiatric Disorders Research, and MIT's Quest for Intelligence.

References

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage*, *26*(3), 839–851.

Assem, M., Glasser, M. F., Van Essen, D. C., & Duncan, J. (2020). A domain-general cognitive core defined in multimodally parcellated human cortex. *Cerebral Cortex*, *30*(8), 4361–4380.

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, *108*(11), 4429–4434. https://doi.org/10.1073/pnas.1014818108

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

Blank, I. A., & Fedorenko, E. (2017). Domain-General Brain Regions Do Not Track Linguistic Input as Closely as Language-Selective Regions. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(41), 9999–10011. https://doi.org/10.1523/JNEUROSCI.3642-16.2017

Blank, I. A., Kanwisher, N., & Fedorenko, E. (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, *112*(5), 1105–1118. https://doi.org/10.1152/jn.00884.2013

Braga, R. M., DiNicola, L. M., Becker, H. C., & Buckner, R. L. (2020). Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *Journal of Neurophysiology*, *124*(5), 1415–1448. https://doi.org/10.1152/jn.00753.2019

Casto, C., Poliak, M., Tuckute, G., Small, H., Sherlock, P., Wolna, A., Lipkin, B., D'Mello, A. M., & Fedorenko, E. (2025). *The cerebellar components of the human language network* (p. 2025.04.14.645351). bioRxiv. https://doi.org/10.1101/2025.04.14.645351

Cohen, L., Salondy, P., Pallier, C., & Dehaene, S. (2021). How does inattention affect written and spoken language processing? *Cortex*, *138*, 212–227.

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, *39*(39), 7722–7736.

Diachek, E., Blank, I. A., Siegelman, M., Affourtit, J., & Fedorenko, E. (2020). The domain-general multiple demand (MD) network does not support core aspects of language comprehension: A large-scale fMRI investigation. *Journal of Neuroscience*, *40*(23), 4536–4550.

Dice, L. R. (1945). Measures of the amount of ecologic association between species.

Ecology, *26*(3), 297–302.
Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science*, *293*(5539), 2470–2473. https://doi.org/10.1126/science.1063414
Du, J., DiNicola, L. M., Angeli, P. A., Saadon-Grosman, N., Sun, W., Kaiser, S., Ladopoulou, J., Xue, A., Yeo, B. T. T., Eldaief, M. C., & Buckner, R. L. (2024). Organization of the human cerebral cortex estimated within individuals: Networks, global topography, and function. *Journal of Neurophysiology*, *131*(6), 1014–1082. https://doi.org/10.1152/jn.00308.2023
Du, J., Tripathi, V., Elliott, M. L., Ladopoulou, J., Sun, W., Eldaief, M. C., & Buckner, R. L. (2025). Within-individual precision mapping of brain networks exclusively using task data. *Neuron*, *0*(0). https://doi.org/10.1016/j.neuron.2025.08.029
Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004
Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*(10), 475–483.
Elhilali, M., Fritz, J. B., Chi, T.-S., & Shamma, S. A. (2007). Auditory cortical receptive fields: Stable entities with plastic abilities. *Journal of Neuroscience*, *27*(39), 10372–10382.
Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402
Fedorenko, E. (2021). The early origins and the growing popularity of the individual-subject analytic approach in human neuroscience. *Current Opinion in Behavioral Sciences*, *40*, 105–112.
Fedorenko, E., & Blank, I. A. (2020). Broca's Area Is Not a Natural Kind. *Trends in Cognitive Sciences*, *24*(4), 270–284. https://doi.org/10.1016/j.tics.2020.01.001
Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *110*(41), 16616–16621. https://doi.org/10.1073/pnas.1315235110
Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010
Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, *25*(5), 289–312.
Francis, W. N., & Kučera, H. (1964). *The Brown Corpus: A Standard Corpus of Present-Day Edited American English for Use with Digital Computers* [Corpus]. Department of Linguistics, Brown University.

Friston, Karl. J., Ashburner, J., Frith, C. D., Poline, J. -B., Heather, J. D., & Frackowiak, R. S. J. (1995). Spatial registration and normalization of images. *Human Brain Mapping*, *3*(3), 165–189. https://doi.org/10.1002/hbm.460030303

Giglio, L., Ostarek, M., Weber, K., & Hagoort, P. (2022). Commonalities and Asymmetries in the Neurobiological Infrastructure for Language Production and Comprehension. *Cerebral Cortex (New York, N.Y.: 1991)*, *32*(7), 1405–1418. https://doi.org/10.1093/cercor/bhab287

Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences*, *111*(10). https://doi.org/10.1073/pnas.1312567111

Hiersche, K. J., Schettini, E., Li, J., & Saygin, Z. M. (2024). Functional dissociation of the language network and other cognition in early childhood. *Human Brain Mapping*, *45*(9), e26757. https://doi.org/10.1002/hbm.26757

Hu, J., Small, H., Kean, H., Takahashi, A., Zekelman, L., Kleinman, D., Ryan, E., Nieto-Castañón, A., Ferreira, V., & Fedorenko, E. (2023). Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, *33*(8), 4384–4404.

Issa, E. B., & Wang, X. (2008). Sensory responses during sleep in primate primary and secondary auditory cortex. *Journal of Neuroscience*, *28*(53), 14467–14480.

Ivanova, A. A., Mineroff, Z., Zimmerer, V., Kanwisher, N., Varley, R., & Fedorenko, E. (2021). The Language Network Is Recruited but Not Required for Nonverbal Event Semantics. *Neurobiology of Language*, *2*(2), 176–201. https://doi.org/10.1162/nol_a_00030

Jäncke, L., Mirzazade, S., & Shah, N. J. (1999). Attention modulates activity in the primary and the secondary auditory cortex: A functional magnetic resonance imaging study in human subjects. *Neuroscience Letters*, *266*(2), 125–128.

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *Neuroimage*, *60*(4), 2357–2364.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Kean, H. H., Fung, A., Pramod, R. T., Chomik-Morales, J., Kanwisher, N., & Fedorenko, E. (2025). Intuitive physical reasoning is not mediated by linguistic nor exclusively domain-general abstract representations. *Neuropsychologia*, 109125.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633.

https://doi.org/10.3758/BRM.42.3.627
Koida, K., & Komatsu, H. (2007). Effects of task demands on the responses of color-selective neurons in the inferior temporal cortex. *Nature Neuroscience*, *10*(1), 108–116.
Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.
Lee, J. J., Scott, T. L., & Perrachione, T. K. (2024). Efficient functional localization of language regions in the brain. *NeuroImage*, *285*, 120489. https://doi.org/10.1016/j.neuroimage.2023.120489
Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., Jouravlev, O., Rakocevic, L., Pritchett, B., & Siegelman, M. (2022). Probabilistic atlas for the language network based on precision fMRI data from> 800 individuals. *Scientific Data*, *9*(1), 529.
MacGregor, L. J., Gilbert, R. A., Balewski, Z., Mitchell, D. J., Erzinçlioğlu, S. W., Rodd, J. M., Duncan, J., Fedorenko, E., & Davis, M. H. (2022). Causal Contributions of the Domain-General (Multiple Demand) and the Language-Selective Brain Networks to Perceptual and Semantic Challenges in Speech Comprehension. *Neurobiology of Language*, *3*(4), 665–698. https://doi.org/10.1162/nol_a_00081
Mahowald, K., & Fedorenko, E. (2016). 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. https://doi.org/10.1016/j.neuroimage.2016.05.073
Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). *Dissociating language and thought in large language models* (No. arXiv:2301.06627). arXiv. http://arxiv.org/abs/2301.06627
Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), 1014–1019.
Marvi, A. I., Hutchinson, S., Fedorenko, E., Saxe, R. R., Kamps, F. S., Regev, T. I., Chen, E. M., & Kanwisher, N. G. (2025). An efficient multifunction fMRI localizer for high-level visual, auditory, and cognitive regions in humans. *Imaging Neuroscience*, *3*, IMAG.a.905. https://doi.org/10.1162/IMAG.a.905
McKee, J. L., Riesenhuber, M., Miller, E. K., & Freedman, D. J. (2014). Task dependence of visual and category representations in prefrontal and inferior temporal cortices. *Journal of Neuroscience*, *34*(48), 16065–16075.
Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared Language:

Overlap and Segregation of the Neuronal Infrastructure for Speaking and Listening Revealed by Functional MRI. *Psychological Science*, *22*(9), 1173–1182. https://doi.org/10.1177/0956797611418347

Mineroff, Z., Blank, I. A., Mahowald, K., & Fedorenko, E. (2018). A robust dissociation among the language, multiple demand, and default mode networks: Evidence from inter-region correlations in effect size. *Neuropsychologia*, *119*, 501–511.

Neville, H. J., Bavelier, D., Corina, D., Rauschecker, J., Karni, A., Lalwani, A., Braun, A., Clark, V., Jezzard, P., & Turner, R. (1998). Cerebral organization for language in deaf and hearing subjects: Biological constraints and effects of experience. *Proceedings of the National Academy of Sciences*, *95*(3), 922–929. https://doi.org/10.1073/pnas.95.3.922

Newport, E. L., Seydell-Greenwald, A., Landau, B., Turkeltaub, P. E., Chambers, C. E., Martin, K. C., Rennert, R., Giannetti, M., Dromerick, A. W., Ichord, R. N., Carpenter, J. L., Berl, M. M., & Gaillard, W. D. (2022). Language and developmental plasticity after perinatal stroke. *Proceedings of the National Academy of Sciences*, *119*(42), e2207293119. https://doi.org/10.1073/pnas.2207293119

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Olson, H. A., Johnson, K. T., Nishith, S., Frosch, I. R., Gabrieli, J. D., & D'Mello, A. M. (2024). Personalized neuroimaging reveals the impact of children's interests on language processing in the brain. *Imaging Neuroscience*, *2*, 1–14.

Olulade, O. A., Seydell-Greenwald, A., Chambers, C. E., Turkeltaub, P. E., Dromerick, A. W., Berl, M. M., Gaillard, W. D., & Newport, E. L. (2020). The neural basis of language development: Changes in lateralization over age. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(38), 23477–23483. https://doi.org/10.1073/pnas.1905590117

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527. https://doi.org/10.1073/pnas.1018711108

Poghosyan, V., & Ioannides, A. A. (2008). Attention modulates earliest responses in the primary auditory and visual cortices. *Neuron*, *58*(5), 802–813.

Polimeni, J. R., & Lewis, L. D. (2021). Imaging faster neural dynamics with fast fMRI: A need for updated models of the hemodynamic response. *Progress in Neurobiology*, *207*, 102174. https://doi.org/10.1016/j.pneurobio.2021.102174

Quillen, I. A., Yen, M., & Wilson, S. M. (2021). Distinct Neural Correlates of Linguistic and Non-Linguistic Demand. *Neurobiology of Language*, *2*(2), 202–225. https://doi.org/10.1162/nol_a_00031

Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and invariant neural responses to spoken and written narratives. *Journal of Neuroscience*,

33(40), 15978–15988.

Richardson, H., Koster-Hale, J., Caselli, N., Magid, R., Benedict, R., Olson, H., Pyers, J., & Saxe, R. (2020). Reduced neural selectivity for mental states in deaf children with delayed exposure to sign language. *Nature Communications*, *11*(1), 3246. https://doi.org/10.1038/s41467-020-17004-y

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage*, *30*(4), 1088–1096; discussion 1097-1099. https://doi.org/10.1016/j.neuroimage.2005.12.062

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*(3), 167–176. https://doi.org/10.1080/17588928.2016.1201466

Sela, Y., Vyazovskiy, V. V., Cirelli, C., Tononi, G., & Nir, Y. (2016). Responses in rat core auditory cortex are preserved during sleep spindle oscillations. *Sleep*, *39*(5), 1069–1082.

Shain, C., Blank, I. A., Fedorenko, E., Gibson, E., & Schuler, W. (2022). Robust Effects of Working Memory Demand during Naturalistic Language Comprehension in Language-Selective Cortex. *Journal of Neuroscience*, *42*(39), 7412–7430. https://doi.org/10.1523/JNEUROSCI.1894-21.2022

Shain, C., & Fedorenko, E. (2025). A language network in the individualized functional connectomes of over 1,000 human brains doing arbitrary tasks. *bioRxiv*, 2025–03.

Shain, C., Kean, H., Casto, C., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., & Fedorenko, E. (2024). Distributed sensitivity to syntax and semantics throughout the language network. *Journal of Cognitive Neuroscience*, *36*(7), 1427–1471.

Shashidhara, S., Assem, M., Glasser, M. F., & Duncan, J. (2024). Task and stimulus coding in the multiple-demand network. *Cerebral Cortex (New York, N.Y.: 1991)*, *34*(7), bhae278. https://doi.org/10.1093/cercor/bhae278

Shashidhara, S., Mitchell, D. J., Erez, Y., & Duncan, J. (2019). Progressive recruitment of the frontoparietal multiple-demand system with increased task complexity, time pressure, and reward. *Journal of Cognitive Neuroscience*, *31*(11), 1617–1630.

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*(6869), 318–320.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, *111*(43).

https://doi.org/10.1073/pnas.1323812111
Somers, D. C., Dale, A. M., Seiffert, A. E., & Tootell, R. B. H. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proceedings of the National Academy of Sciences*, *96*(4), 1663–1668. https://doi.org/10.1073/pnas.96.4.1663
Sueoka, Y., Paunov, A., Tanner, A., Blank, I. A., Ivanova, A., & Fedorenko, E. (2024). The Language Network Reliably “Tracks” Naturalistic Meaningful Nonverbal Stimuli. *Neurobiology of Language*, *5*(2), 385–408.
Terhune-Cotter, B. P. (2025). *Exploring the neural network for sign language comprehension across deaf signers* [UC San Diego]. https://escholarship.org/uc/item/9698p29n
Tuckute, G., Lee, E. J., Sathe, A., & Fedorenko, E. (2024). *A 3.5-minute-long reading-based fMRI localizer for the language network* (p. 2024.07.02.601683). bioRxiv. https://doi.org/10.1101/2024.07.02.601683
Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C., & Dehaene, S. (2012). A temporal bottleneck in the language comprehension network. *Journal of Neuroscience*, *32*(26), 9089–9102.
Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., Smith, N., Gibson, E., & Fedorenko, E. (2021). Incremental Language Comprehension Difficulty Predicts Activity in the Language Network but Not the Multiple Demand Network. *Cerebral Cortex*, *31*(9), 4006–4023. https://doi.org/10.1093/cercor/bhab065
Wilson, S. M., Bautista, A., Yen, M., Lauderdale, S., & Eriksson, D. K. (2017). Validity and reliability of four language mapping paradigms. *NeuroImage: Clinical*, *16*, 399–408. https://doi.org/10.1016/j.nicl.2016.03.015
Wolna, A., Szewczyk, J., Diaz, M., Domagalik, A., Szwed, M., & Wodniecka, Z. (2024). Tracking Components of Bilingual Language Control in Speech Production: An fMRI Study Using Functional Localizers. *Neurobiology of Language*, *5*(2), 315–340. https://doi.org/10.1162/nol_a_00128
Woolgar, A., Hampshire, A., Thompson, R., & Duncan, J. (2011). Adaptive Coding of Task-Relevant Information in Human Frontoparietal Cortex. *The Journal of Neuroscience*, *31*(41), 14592–14599. https://doi.org/10.1523/JNEUROSCI.2616-11.2011
Zanto, T. P., & Gazzaley, A. (2013). Fronto-parietal network: Flexible hub of cognitive control. *Trends in Cognitive Sciences*, *17*(12), 602–603.