Cerebral Cortex, 2024, **34**, bhae077
https://doi.org/10.1093/cercor/bhae077
Advance access publication date 17 March 2024
Original Article

High-level language brain regions process sublexical regularities

Tamar I. Regev[1,2,*], Hee So Kim[1,2], Xuanyi Chen[1,2,3], Josef Affourtit[1,2], Abigail E. Schipper[1], Leon Bergen[4], Kyle Mahowald[5,†], Evelina Fedorenko[1,2,6,†]

[1]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, United States
[2]McGovern Institute for Brain Research, MIT, Cambridge, MA 02139, United States
[3]Department of Cognitive Sciences, Rice University, Houston, TX 77005, United States
[4]Department of Linguistics, University of California San Diego, San Diego CA 92093, United States
[5]Department of Linguistics, University of Texas at Austin, Austin, TX 78712, United States
[6]The Harvard Program in Speech and Hearing Bioscience and Technology, Boston, MA 02115, United States
*Corresponding author: Department of Brain and Cognitive Sciences, MIT, 43 Vassar Street, Room 46-4127B, Cambridge, MA 02139, United States.
Email: tamarr@mit.edu
†Kyle Mahowald and Evelina Fedorenko Co-senior authors.

A network of left frontal and temporal brain regions supports language processing. This "core" language network stores our knowledge of words and constructions as well as constraints on how those combine to form sentences. However, our linguistic knowledge additionally includes information about phonemes and how they combine to form phonemic clusters, syllables, and words. Are phoneme combinatorics also represented in these language regions? Across five functional magnetic resonance imaging experiments, we investigated the sensitivity of high-level language processing brain regions to sublexical linguistic regularities by examining responses to diverse nonwords—sequences of phonemes that do not constitute real words (e.g. punes, silory, flope). We establish robust responses in the language network to visually (experiment 1a, $n = 605$) and auditorily (experiments 1b, $n = 12$, and 1c, $n = 13$) presented nonwords. In experiment 2 ($n = 16$), we find stronger responses to nonwords that are more well-formed, i.e. obey the phoneme-combinatorial constraints of English. Finally, in experiment 3 ($n = 14$), we provide suggestive evidence that the responses in experiments 1 and 2 are not due to the activation of real words that share some phonology with the nonwords. The results suggest that sublexical regularities are stored and processed within the same fronto-temporal network that supports lexical and syntactic processes.

Key words: language; brain; fMRI; phonology; sublexical.

Introduction

Languages contain rich statistical patterns across a range of information scales—from inter-word dependencies, to meanings of individual words and morphemes, to patterns of sounds within words—but whether linguistic information at different scales is represented and processed by overlapping or distinct cognitive and neural mechanisms remains debated. Traditionally, a distinction has been drawn between "high-level" linguistic processes, such as syntax and lexical semantics, and phonological processing, which was considered lower level and thus assumed to rely on distinct cognitive and neural machinery (e.g. Chomsky 1965, 1995; Chomsky and Halle 1965; Bromberger and Halle, 1989; Pinker 1991; Heinz and Idsardi, 2011, 2013; Berwick and Chomsky, 2016; see Matchin (2018) for recent implementation-level claims about the separation between phonological and higher-level processes). However, some linguistic theories have suggested a more integrated view of language processing, where the boundaries between our processing of the sentence structure, word meanings, and sublexical sound patterns are less sharp (e.g. Gaskell and Marslen-Wilson 1997; Bybee 1999, 2013; Goldberg 2003; Jackendoff 2007; Huettig et al. 2020; Jackendoff and Audring 2020).

In support of this integrated view of language processing, corpus investigations across diverse languages have revealed strong relationships between sound patterns and other aspects of language. For instance, more frequent words tend to be more *phonotactically regular*, i.e. obeying the phoneme-combinatorial constraints of the language (e.g. Zipf 1936; Landauer and Streeter 1973; Frauenfelder et al. 1993; Mahowald et al. 2018; Pimentel et al. 2020), phonological clustering may be one organizing principle of the lexicon (e.g. Dautriche et al. 2017), and some sounds/sound patterns appear to be associated with aspects of meaning (e.g. Iwasaki et al. 2007; Monaghan et al. 2014; Larsson 2015; Blasi et al. 2016; Winter et al. 2017; Sidhu and Pexman 2018; Pimentel et al. 2019; Vinson et al. 2021). Further, sound patterns can differentiate syntactic categories, like nouns and verbs (e.g. Kelly 1992; Albright 2008; Arciuli and Monaghan 2009; Arciuli et al. 2012). These links between sound patterns and other aspects of linguistic structure and meaning may be particularly important for language acquisition as linguistic input is initially perceived as a meaningless sequence of sounds that the language system attempts to interpret. Indeed, early word learning is facilitated by sound–meaning associations or iconicity (Perry et al. 2018) and by knowledge of phonotactic regularities (Storkel 2001; Coady and Aslin 2004; Dautriche et al. 2015; de Carvalho et al. 2016; Jones et al. 2021); this knowledge continues to facilitate lexical access in adulthood (e.g. Vitevitch et al. 1999; Vitevitch and Luce 1999; Luce and Large 2001).

Does strong integration between sound patterns and lexical or syntactic features mean that—at the implementation level—the

Received: August 12, 2023. **Revised:** February 5, 2024. **Accepted:** February 7, 2024

system that processes words and sentences (i.e. supports computations related to lexical access, syntactic structure building, and semantic composition) also processes sublexical sound patterns? Past neuroscience research has not provided a clear answer. Prior neuroimaging investigations have reported effects for phonological manipulations in diverse left-hemisphere (or bilateral) brain areas, including superior temporal gyrus (e.g. Paulesu et al. 1993; Price et al. 1997; Okada and Hickok 2006; Graves et al. 2007, 2008; DeWitt and Rauschecker 2012; Gow and Olson 2015; Lopopolo et al. 2017; Scott and Perrachione 2019), supramarginal gyrus (e.g. Paulesu et al. 1993; Celsis et al. 1999; Church et al. 2011; Weiss et al. 2018; Yen et al. 2019), and inferior frontal cortex (e.g. Paulesu et al. 1993; Demonet et al. 1994; Poldrack et al. 1999; Burton 2001; Myers et al. 2009; Vaden et al. 2011; Okada et al. 2017; Xie and Myers 2018). Similarly, lesions in these different brain areas (e.g. Geva et al. 2011; Pillay et al. 2014; Kries et al. 2023), as well as their interruption by electric/magnetic stimulation (e.g. Devlin et al. 2003; Boatman 2004; Hartwigsen et al. 2016), have been shown to lead to impairments on phonological tasks, like rhyme judgments, nonword repetition, or phoneme identification.

Some of the brain areas implicated in phonological processing appear to overlap with the "core" language network—a set of left-lateralized frontal and temporal areas that selectively respond to linguistic input, visual or auditory (e.g. Fedorenko et al. 2011; Monti et al. 2012) and support the processing of word forms and meanings and combinatorial syntactic and semantic processes (e.g. Bozic et al. 2010; Fedorenko et al. 2010, 2020; Bautista and Wilson 2016). However, inferences about shared vs. distinct neural mechanisms based on the similarity of gross anatomical locations across studies are problematic (e.g. Poldrack 2006; Fedorenko 2021). Furthermore, most past studies of phonological processing have employed tasks that differ in their computational demands from those of naturalistic language processing, where the goal is to simply extract meaning from the linguistic input. Some studies have required (overt or covert) speech production and may have therefore recruited the speech articulation system (e.g. Bohland and Guenther 2006; Basilakos et al. 2017), and others have used tasks with executive demands (e.g. rhyme judgments) and may have therefore recruited domain-general executive resources (see, e.g. Diachek et al. 2020; Quillen et al. 2021 for evidence that the executive control system gets engaged when language comprehension is accompanied by extraneous tasks).

To provide a clearer answer about whether the system that supports lexical and word-combinatorial processing is sensitive to sublexical sound patterns, we functionally defined the language network using an established language "localizer" task (Fedorenko et al. 2010) and then examined these brain regions' responses to nonwords—sequences of phonemes that do not constitute real English words—during relatively naturalistic reading/listening across five fMRI experiments. It is important to note that although we define the language regions as regions that respond more strongly during sentence processing compared to the processing of nonwords (or similar control conditions; Methods), this definition does not entail that the response to nonwords would be negligible. In fact, we have previously observed that the response to nonwords in these language regions is consistently above a low-level fixation baseline (Fedorenko et al. 2010; Blank et al. 2016; Fedorenko and Blank 2020). We here formally investigate this effect.

To foreshadow the key findings, visually and auditorily presented nonwords elicited robust responses across the language network despite their lack of meaning and lack of ability to combine into larger units like phrases. Further, nonwords that were more well-formed elicited stronger responses than less well-formed ones, which suggests that the language network represents and processes phoneme-combinatorial regularities. We further provide suggestive evidence that the response to nonwords in the language network is not merely due to the activation of representations of real words that share some phonology with the nonwords, thus strengthening the claim that sublexical meaningless units are processed by the same system that processes words, phrases, and sentences.

Materials and methods

Participants

In total, 620 individuals (age 18 to 71 mean 24.9 + −7.3; 358 [57.7%] females) from the Cambridge/Boston, MA community participated for payment across five fMRI experiments ($n = 605$ in experiment 1a, $n = 12$ in experiment 1b, $n = 13$ in experiment 1c, $n = 16$ in experiment 2, and $n = 14$ in experiment 3, for a total of 660 scanning sessions; Table 1). For experiment 1a, we leveraged a large dataset that was collected in our lab across 10+ years (Lipkin et al. 2022). Forty participants overlapped between experiment 1a and other experiments (12, 14, and 14 with experiments 1c, 2, and 3, respectively; Table 1) and 4 participants overlapped between experiments 2 and 3. 558 participants (~90%, see Table 1 for numbers per experiment) were right-handed, as determined by the Edinburgh handedness inventory (Oldfield, 1971), or self-report; the remaining participants were either left-handed ($n = 40$), ambidextrous ($n = 14$), or missing handedness information ($n = 8$; see Willems et al. 2014 for arguments for including left-handers in cognitive neuroscience research). All participants were native English speakers, and all gave written informed consent to participate in our experiments in accordance with the requirements of Massachusetts Institute of Technology (MIT)'s Committee on the Use of Humans as Experimental Subjects (COUHES), which approved the study.

Design, materials, and procedure

All experiments—overview

In all experiments, we examined responses to nonwords—meaningless sound/letter strings—in the high-level language system. Therefore, in all experiments each participant completed an fMRI reading–based *language network "localizer" task* based on contrasting fMRI responses between reading sentences and reading nonword sequences, as detailed below. This localizer was previously shown to be robust to modality—the same brain regions are found in a reading-based or listening-based version of the localizer (e.g. Fedorenko et al. 2010, Scott et al. 2017, Chen et al. 2021, Malik-Moraleda et al. 2022). Participants also completed one or more tasks, including the *critical experimental task* (the main task used for each experiment in this study) and, in most cases, other tasks for unrelated studies. The total session duration was typically around 2 h.

The purpose of experiments 1a, 1b, and 1c was to examine the general robustness of responses to nonwords within the language network, across the visual (reading) and auditory (listening) modalities. The nonwords in experiments 1a, 1b, and 1c, as well as in the language localizer, were all constructed to meet the phoneme-combinatorial constraints of English (and thus to sound relatively well-formed) using slightly different methods, as detailed below for each experiment. The purpose of experiments 2 and 3 was to examine how phonological characteristics of the nonwords affected neural responses. In experiment 2, we manipulated nonword well-formedness (which correlates with

Table 1. Details of participants in all experiments.

Experiment	1a	1b	1c	2	3
Participants	605	12	13	16	14
Females	349	7	6	11	11
Age (mean, std) (years)	24.9 (7.3)	23.2 (4)	24.7 (6.7)	22.2 (7.1)	25.1 (7.6)
Left-handed (ambidextrous)	36 (13)	0 (0)	1 (0)	2 (0)	1 (2)
Participants overlapping with experiment:					
1a		0	12	14	14
1b			0	0	0
1c				0	0
2					4
3					

phonotactic probability), and in experiment 3 we manipulated the neighborhood density of nonwords (i.e. the number of real words that are phonologically similar to the nonword; Vitevitch et al. 1999), as detailed below.

Reading-based language network localizer

This task was originally designed to elicit robust responses in the high-level language network, as described in detail in Fedorenko et al. (2010) and subsequent studies from the Fedorenko lab (and is available for download from https://evlab.mit.edu/funcloc/). In this task, participants read sentences and lists of unconnected pronounceable nonwords in a blocked design and were asked to press a button at the end of each trial, when a special symbol appeared, to maintain alertness. The words or nonwords appeared on the screen one at a time. The vast majority of participants (605 out of 620) performed a version of the localizer where the nonwords were created using the Wuggy software (Keuleers and Brysbaert 2010), to match their phonotactic properties to those of the words used in the sentence condition. See further details in Supplementary Information Section 1 (*Timing and details of stimulus presentation*). The *Sentences* > *Nonwords* contrast targets brain regions that support high-level language comprehension, including lexico-semantic and combinatorial (syntactic and compositional semantic) processes (e.g. Fedorenko et al. 2010, 2020; Fedorenko et al. 2012b; Blank et al. 2016), and has been shown to be robust to changes in modality (visual/auditory), materials, task, timing parameters, and other aspects of the procedure (e.g. Fedorenko et al. 2010; Fedorenko 2014; Mahowald and Fedorenko 2016; Scott et al. 2017; Diachek et al. 2020). As such, this specific contrast is standardly used in our lab as a language localizer contrast, but many similar contrasts work equally well (see Fedorenko et al. in press, for a review).

Experiment 1a (passive reading of lists of nonwords from the language localizer)

To examine the robustness of responses to visually presented nonwords in the language regions, we used the nonwords condition from the reading-based language network localizer. Response magnitudes were estimated using cross-validation across experimental runs to ensure that the data used for the localization of the language regions were independent from the data used to estimate the responses to nonwords in this critical task (e.g. Kriegeskorte et al. 2009). The cross-validation was performed in the following way: First, run 1 was used to define the regions of interest and run 2 to estimate the responses (each participant performed 2 runs of the task); then, run 2 was used to define the regions and run 1 to estimate the responses; finally, the estimates were averaged across the two runs to derive a single estimate per participant per region. As noted above (Reading-based language network localizer), the nonwords were accompanied by a simple button-press task to maintain alertness. Behavioral responses for this and all other experiments are summarized in a supplementary table available at the Open Science Framework (OSF) platform (https://osf.io/6c2y7/). Example items are shown in Fig. 1.

Experiment 1b (listening to lists of nonwords followed by a memory probe)

To examine the robustness of responses to auditorily presented nonwords in the language regions, we used the nonwords condition from an auditory language experiment that was published previously (experiment 3 in Fedorenko et al. 2010). Participants listened to recordings of lists of nonwords (and materials from three other conditions that are not relevant to the current study) in a blocked design and, at the end of each trial, judged whether a probe word/nonword appeared in the trial. The nonwords were constructed by recombining the syllables that comprised the words in the real-word conditions of the experiment (to preserve phonotactic well-formedness) and were recorded by a female native English speaker (see Fedorenko et al. 2013b for a detailed acoustic analysis of these materials). Example items are shown in Fig. 1. See further details in Supplementary Information Section 1 (*Timing and details of stimulus presentation*).

Experiment 1c (passive listening to lists of nonwords)

To replicate and generalize the results from experiment 1b, we used a nonwords condition from another auditory experiment (experiment 4 in Chen et al. 2023). Participants passively listened to recordings of lists of nonwords (and materials from several other conditions that are not relevant to the current study) in a blocked design. The nonword lists were constructed by taking a set of sentences and replacing each word with a nonword that has a similar phonological structure (taking into account consonant–vowel structure, consonant class, vowel class, and rhythmicity) but that does not have any meaning. These "nonword sentences" were recorded by a female and a male native English speaker. In the experiment, half of the trials came from the female speaker, and the other half from the male speaker. Example items are shown in Fig. 1, and the full list of materials is available at OSF (https://osf.io/6c2y7/). See further details in Supplementary Information Section 1 (*Timing and details of stimulus presentation*).

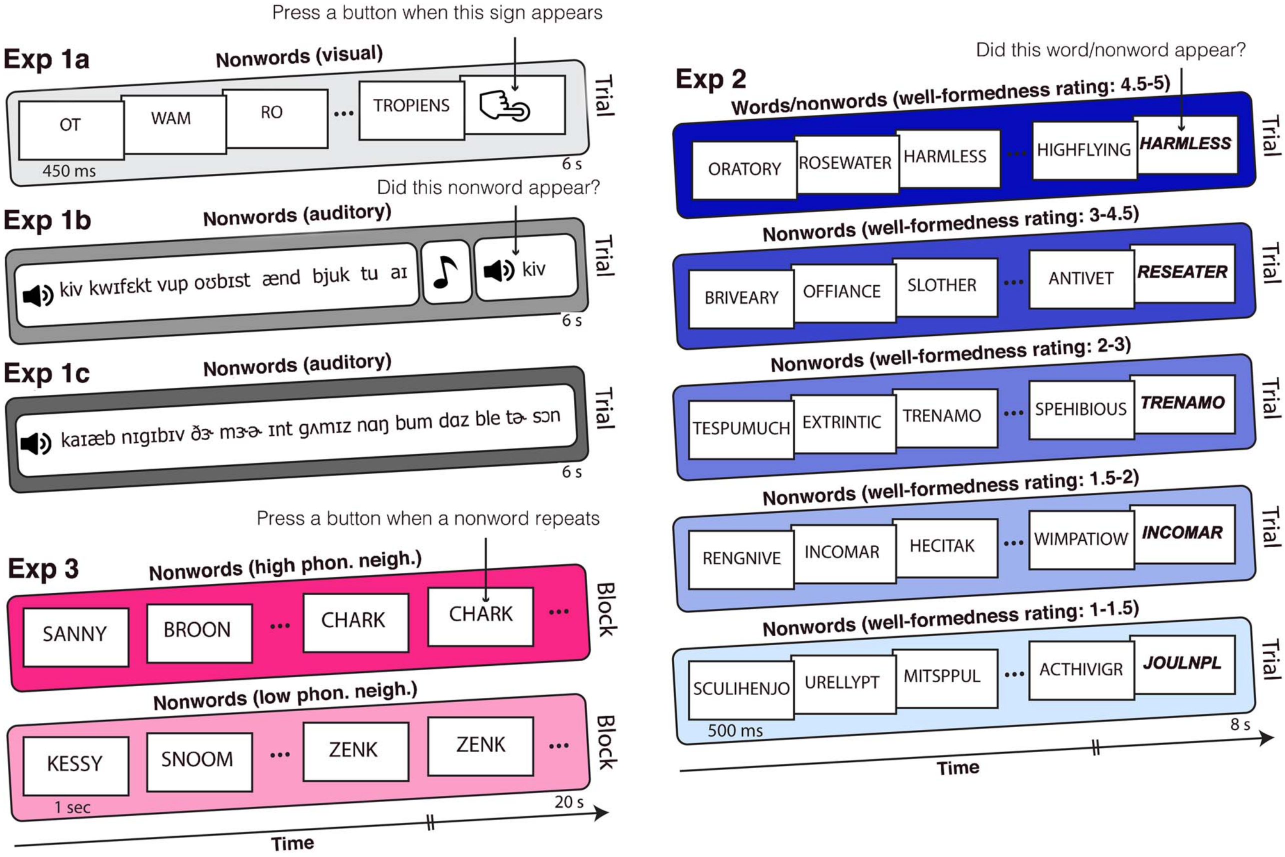

Fig. 1. Procedure and example stimuli for all experiments. Color-filled rounded rectangles represent a typical block or trial in a specific condition of each experiment. The color codes match those used in Fig. 2. Experiment numbers and conditions are indicated above each rectangle. Left to right, top to bottom (see Methods for further detail): *Exp 1a*—passive reading of lists of nonwords from the language localizer. *Exp 1b*—listening to lists of nonwords followed by a memory probe. *Exp 1c*—passive listening to lists of nonwords. *Exp 2*—reading of lists of nonwords parametrically varying in well-formedness, followed by a memory probe. *Exp 3*—reading of lists of nonwords with a low or high phonological neighborhood, accompanied by repetition detection (the experiments are not ordered due to their ordinal numbers to preserve space in the figure).

Experiment 2 (reading lists of nonwords—that vary in their well-formedness—followed by a memory probe)

To test whether more well-formed nonwords would elicit a stronger response in the language regions, participants read lists of real words and nonwords (and materials from five other conditions that are not relevant to the current study) in a blocked design. The nonwords were created from real words via one or multiple letter replacements, as detailed below. The original words and different resulting versions of nonwords were grouped into five conditions based on well-formedness ratings, which were obtained in a behavioral norming study conducted online, with independent participants, as described below. Example items are shown in Fig. 1, and all items are available at OSF (https://osf.io/6c2y7/). See further details in Supplementary Information Section 1 (*Timing and details of stimulus presentation*).

Construction and norming of the materials

To create the nonwords, a large set of real trisyllabic English words ($n = 20{,}695$) was first identified. For each word, 14 versions of nonwords were created by iteratively replacing random letters with other letters, while ensuring that the local trigram context (the letter preceding the critical letter, the critical replaced letter, and the letter following it) is attested in English (i.e. appears in at least one real word). For example, consider the word "BLACKBERRY"; the letter C could be replaced with the letter R because the string "ARK" is attested (e.g. BARK), or with the letter L because the string "ALK" is attested (e.g. ALKALINE), but not with the letter X because the string "AXK" is not attested. This replacement process was repeated on the resulting nonword (e.g. BLARKBERRY in this example, if R was used as the replacement for C) using the same constraints, up to 14 times total. This procedure resulted in a set of 310,425 words and nonwords including the original words and all the resulting nonwords from the 14 letter-replacement iterations done on each word. A subset of these materials ($n = 900$, sampled ~equally from the 15 "levels" of degradation, i.e. number of replaced letters, between 0 and 14) were presented to participants online via Amazon.com's Mechanical Turk platform. Participants were presented with one word/nonword at a time and asked to rate each for how well-formed it was (the exact wording was: "How close is this to being an acceptable English word?"), on a scale from 1 (very unacceptable) to 5 (very acceptable). The words/nonwords were then divided into five sets according to the well-formedness ratings, from least to most well-formed: 1 to 1.5, 1.5 to 2, 2 to 3, 3 to 4.5, and 4.5 to 5. The bin sizes were determined by the distribution of ratings, to balance the number of items within each set (condition). Each set consisted of 180 items, except for the most well-formed set, for which there were only 173 items. The most well-formed set consisted of mostly real words, and the other four sets consisted exclusively of nonwords. Fifteen 12-item strings were created from these materials for each of the five conditions for presentation in a blocked design experiment (for the most well-formed condition, seven of the items were used twice, never within the same string).

Participants in experiment 2 also performed a non-linguistic (spatial working memory) task (Fedorenko et al. 2013a). Data from this task were used in a control analysis, as a comparison to the critical nonword reading task (Methods). In this task, participants viewed a grid within which locations were randomly flashed sequentially (one at a time for a total of four locations in the easy condition and two at a time for a total of eight locations in the hard condition). At the end of the trial, participants had to indicate the locations they just saw by selecting one of two options via a button press, followed by feedback on the correctness of their response. The *Hard* > *Easy* contrast engages the multiple demand (MD) network, which is robustly distinct from the language network (Fedorenko et al. 2013a). See further details in Supplementary Information Section 1 (*Timing and details of stimulus presentation*).

Experiment 3 (reading lists of nonwords with a low or high phonological neighborhood, accompanied by repetition detection)

Previous work has shown that phonotactic regularity of nonwords (which correlates with perceived well-formedness, as manipulated in experiment 2) tends to correlate with phonological neighborhood (defined as the number of real words that are one edit away from the nonword) (e.g. Vitevitch et al. 1999; Vitevitch and Luce 1999), but these two factors can be disentangled (e.g. Luce and Large 2001). Furthermore, behavioral investigations have suggested that experimental manipulations of phonotactic regularities target local phoneme-combinatorial pattern processing (i.e. sublexical processing) whereas manipulations of phonological neighborhood emphasize holistic (lexical) recognition of phoneme strings (Vitevitch and Luce 1999; Luce and Large 2001). We therefore wanted to test whether the results obtained in experiment 2 (stronger responses to more well-formed nonwords) could be due to activation of real words that sound similar to the nonwords (neighbors), instead of perception of the phonological structure of the nonwords themselves. Stronger neural responses to nonwords with more phonological neighbors compared to nonwords with fewer neighbors would support this possibility. Participants read lists of nonwords that were matched on phonotactic probability and other phonological characteristics (as described below) but critically varied in their phonological neighborhood size in a blocked design and were instructed to press a button when a nonword repeated in a row. See further details in Supplementary Information Section 1 (*Timing and details of stimulus presentation*).

Construction of the materials

To construct two sets of nonwords that are matched on phonotactic probability and other phonological characteristics but vary in their phonological neighborhood size, a 3-gram model over phonemes was used, using the generative procedure described in Dautriche et al. (2017). In particular, each phoneme is generated probabilistically, conditioned on the preceding two phonemes. Using this model, a large set of candidate nonwords was sampled without replacement. Then, 80 pairs of nonwords were selected such that they were matched on length in letters and syllables, on the consonant–vowel patterns, and on phonotactic probability, as measured with a pronunciation-based phonotactic ("BLICK") score (e.g. Hayes and Wilson 2008; Hayes 2012) [a two-sample t-test of BLICK scores between the sets: $t(158) = 0.05$, $P = 0.96$], but critically differed maximally in their phonological neighborhood size. Neighborhood size was estimated as the number of real English words that are one edit away from the nonword. For example, phonological neighbors of the nonword "ZAT" include "BAT," "CAT," and "ZAP," among others (although phonological neighborhood size has some limitations as a measure, such as treating all letter positions equally [cf. Marslen-Wilson 1987; Wedel et al. 2019 for evidence of letter-position effects in word recognition], it is standardly used and has been shown to underlie many behavioral effects in word/nonword processing [for a review, see Vitevitch and Luce 2016]). In the high-neighborhood set, each nonword had at least nine neighbors (mean = 11, SD = 2.6), and in the low-neighborhood group, each nonword had at most three neighbors [mean = 1.85, SD = 0.9, two-sample t-test of neighborhood scores between the groups: $t(158) = 28.7$, $P < 0.0001$]. Example items are shown in Fig. 1, and all items are available at https://osf.io/6c2y7/.

Phonotactic probability and neighborhood size of stimuli in experiments 2 and 3

To investigate which features of the nonword stimuli may contribute to neural responses in the language system, we calculated the phonotactic probability and neighborhood size of the stimuli in a unified manner across experiments 2 and 3. We used the English Lexicon Project (https://elexicon.wustl.edu/, Balota et al. 2007) for both measures. This website allows one to submit lists of written nonwords and outputs a series of characteristics calculated based on an English corpus (Balota et al. 2007). The phonotactic probability measure was computed as the mean bigram frequency, which is the sum of bigram counts (where a bigram is a sequence of two letters like ZA and AT in ZAT) for all the local bigrams within a nonword, divided by the number of bigrams. The neighborhood size measure was computed as the number of real words that can be obtained by changing one letter while preserving the identity and positions of the other letters (i.e. Coltheart's N; Coltheart et al. 1977). We chose to use orthography-based measures and not phonology-based measures (as we originally did when designing the materials for experiment 3) because (a) the stimuli were visually presented to the participants and (b) the pronunciation of many English nonwords is inherently ambiguous because of the non-transparency of English spelling (e.g. the nonword KLOUGH could be pronounced to rhyme with *through*, *trough*, or *tough*). However, our results are robust to whether we use orthography- or phonology-based measures (e.g. the correlation between orthographic and phonological neighborhood size measures for the nonwords in experiment 3 is $r = 0.55$, $P < 0.001$). Having obtained the phonotactic probability and neighborhood size measures for the nonwords in experiments 2 and 3, we calculated the average and standard error across all nonwords in each condition (five conditions varying in well-formedness in experiment 2 and two conditions varying in neighborhood size in experiment 3).

fMRI data acquisition

Experiments 1a, 1c, 2, and 3

Whole-brain structural and functional data were collected on a 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 axial slices with 1 mm isotropic voxels (repetition time (TR) = 2,530 ms; echo time (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; the following parameters were used: 31 4.4 mm thick near-axial slices acquired in an interleaved order (with 10% distance factor), with an in-plane resolution of 2.1 mm × 2.1 mm, FoV in the phase encoding anterior to posterior (A > > P) direction 200 mm and matrix size 96 × 96 voxels, TR = 2,000 ms and TE = 30 ms. The first 10 s of each run were excluded to allow for steady state magnetization.

experiment 1b

This experiment had distinct data acquisition parameters because it was conducted at an earlier point in time (2008 to 2009). Whole-brain structural and functional data were collected on the whole-body 3 Tesla Siemens Trio scanner at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1- weighted structural images were collected in 128 axial slices with 1.33 mm isotropic voxels (TR = 2,000 ms, TE = 3.39 ms). Functional BOLD data were acquired in 3.1 × 3.1 × 4 mm voxels (TR = 2,000 ms, TE = 30 ms) in 32 near-axial slices. The first 4 s of each run were excluded to allow for steady-state magnetization.

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 SD above the mean in global BOLD signal change, or framewise displacement values above 0.9 mm, Nieto 2020). 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 and 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 4th-order spline interpolation for the functional data, and 1 mm isotropic voxels and trilinear interpolation for the structural data. Lastly, the functional data were then smoothed spatially using spatial convolution with a 4 mm FWHM Gaussian kernel.

fMRI data first-level modeling

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). Temporal autocorrelations in the BOLD signal timeseries 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 (ReML). In addition to main condition effects, other model parameters in the GLM design included first-order temporal derivatives for each condition, modeling spatial variability in the HRF delays, as well as nuisance regressors controlling for the effect of slow linear drifts, subject-motion parameters, and potential outlier scans on the BOLD signal.

Definition of the language functional regions of interest

For each critical experiment, we first defined a set of language functional regions of interest (fROIs) using an established language localizer (Fedorenko et al. 2010), which identifies a set of brain regions that respond strongly and selectively during language processing and form an integrated functional network, which we refer to as the "language regions" or the "language network" (Fedorenko et al. in press). To define the fROIs, we used a group-constrained, subject-specific (GcSS) approach (Fedorenko et al. 2010), where each individual map for the *Sentences* > *Nonwords* localizer contrast was intersected with a set of five binary left-hemisphere masks. These masks (Fig. 2; available at http://web.mit.edu/evlab//funcloc/#parcels) were derived from a probabilistic activation overlap map for the same contrast in a large set of participants ($n = 220$) using watershed parcellation, as described in Fedorenko et al. (2010) for a smaller set of participants. Within each mask, a participant-specific language fROI was defined as the top 10% of voxels with the highest t-values for the localizer contrast (see Lipkin et al. 2022 for evidence that the language fROIs are similar when defined with a fixed statistical threshold). Effect sizes for the critical tasks were then estimated in the language fROIs by averaging across the voxels within each participant-specific fROI [see also Analyses of the critical tasks (all experiments) below]. For completeness, we also defined (i) homotopic right-hemisphere language fROIs using the same voxel selection procedure within the mirrored versions of the LH masks (see Supplementary Information Section 3, Fig. S1), as well as (ii) bilateral language fROIs in the angular gyrus (see Supplementary Information Section 4, Fig. S2). Both of these sets of areas are activated by the language localizer contrast but have been shown to dissociate from the core frontal and temporal LH language areas (e.g. Shain, Paunov, Chen et al. 2023).

A whole-brain search for areas sensitive to nonword well-formedness (Experiment 2)

In addition to a targeted analysis of the language regions, we searched across the brain for regions that process phoneme-combinatorial regularities and thus exhibit sensitivity to the nonword well-formedness gradient in the critical task in experiment 2. To do so, we used a group-constrained subject-specific (GcSS) analysis, which is more sensitive than a traditional fMRI group analysis given that it takes into account inter-individual variability in the precise locations of functional areas. Using data from experiment 2, we defined a contrast based on the parametric well-formedness manipulation: $\text{GradientW} = -1*\text{W1} - 0.5*\text{W2} + 0*\text{W3} + 0.5*\text{W4} + 1*\text{W5}$, where W1 to W5 are the five conditions in experiment 2 (nonwords that vary in well-formedness from low to high; condition W5 almost exclusively consists of real words). Individual activation maps for this contrast were binarized in the following way: voxels that pass the significance threshold of $P < 0.01$ for the parametric contrast above were denoted as 1 and the rest of the voxels as 0 (note that the use of a relatively liberal threshold at this stage is acceptable because it is only used to create probabilistic maps, and the statistical tests are performed at a later stage with independent data as explained next). Individual binarized activation maps were overlaid to create a probabilistic activation overlap map, which was then parcellated using a watershed algorithm, as described in Fedorenko et al. (2010; a custom Matlab toolbox for doing this is available at https://evlab.mit.edu/funcloc) to

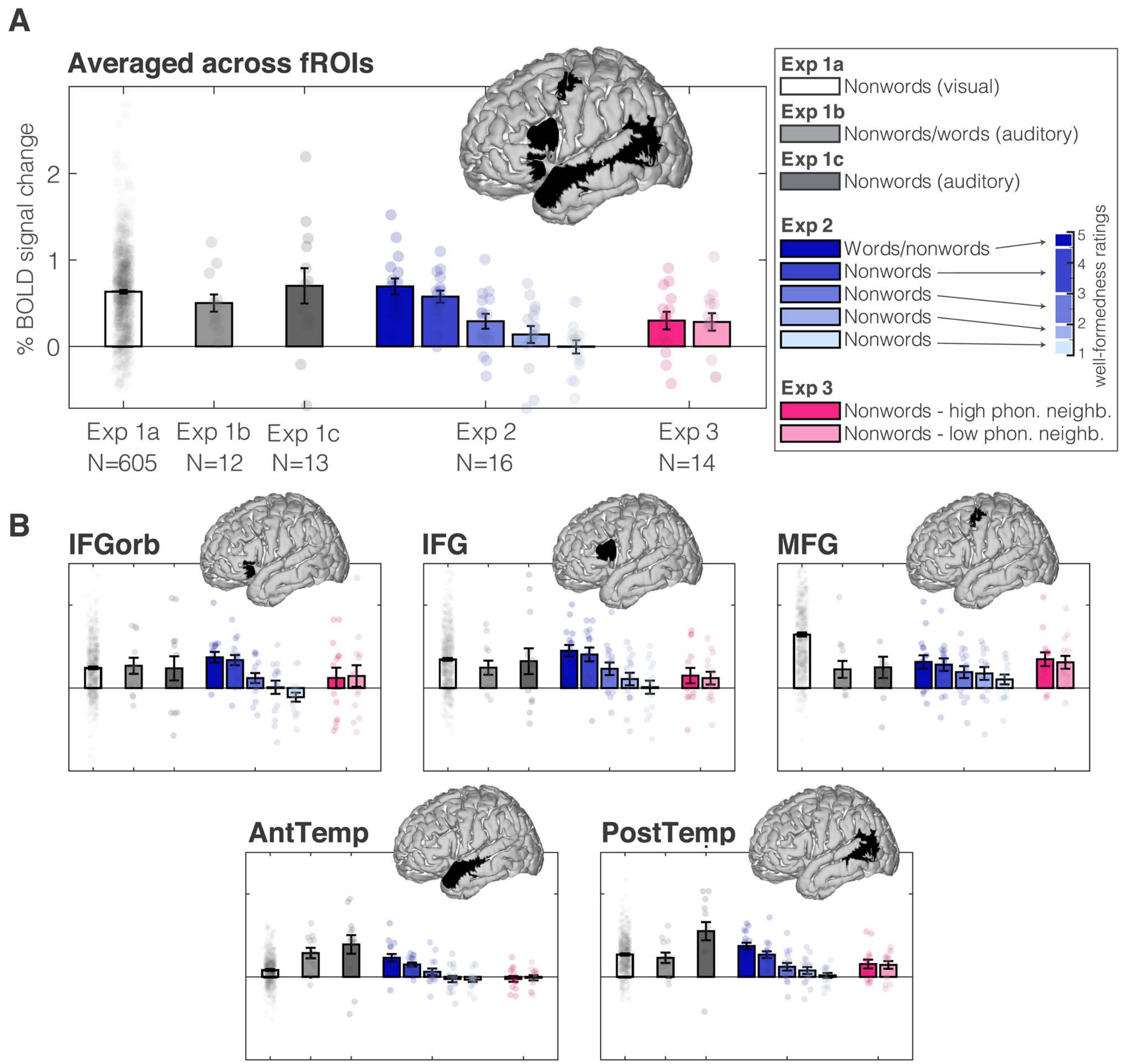

Fig. 2. Responses of the left-hemisphere language network to nonwords in all experiments. Bar graphs show % BOLD signal change relative to a fixation baseline in individually defined language fROIs averaged across participants in each specific experiment (number of participants specified on abscissa). Here and elsewhere, error bars denote standard errors of the mean by participants, and dots are individual participants. The brain images display the masks that were used to define the fROIs; individual fROIs are 10% of most language-responsive voxels within each mask; the effects are estimated using data that are independent from the data used to define the fROIs. A) Responses in all five language fROIs together. B) Responses in each fROI separately. IFGorb, inferior frontal gyrus orbital, IFG, inferior frontal gyrus, MFG, medial frontal gyrus, AntTemp, anterior temporal, PostTemp, posterior temporal (see Fig. S1 for responses in the right-hemisphere homotopic language fROIs, and Fig. S2 for responses in the language fROIs located in the bilateral angular gyri).

identify areas of common activation across participants. Using the resulting masks, we then defined individual fROIs using the top 10% of voxels responding to the contrast above in each individual (similar to how the language fROIs were defined using the group-level language masks and the top 10% voxels responding to *Sentences* > *Nonwords* within the masks in each individual). Then, using an across-runs cross-validation procedure (described above, in Definition of the language functional regions of interest), we estimated the responses within these fROIs to the five conditions (W1 to W5) as well as the *Sentences* and *Nonwords* conditions from the language localizer. We report the results for fROIs that showed a replicable (across runs) well-formedness gradient (W1 < W2 < W3 < W4 < W5) and an above-baseline response to W5 (see Supplementary Information Section 5 for the full set of results).

Comparison of voxel-level activation patterns between the nonword well-formedness contrast and the language localizer contrast (experiment 2)

To quantify the similarity of the activation patterns between the critical contrast in experiment 2 (sensitivity to nonword well-formedness; GradientW contrast, see above) and the language localizer contrast in a way that is not biased by the use of the functional localization approach, we examined voxel-wise correlations in the contrast values (across the brain as well as within the language masks). The correlations were computed for each individual participant ($n = 16$), Fisher-transformed, and then averaged across participants. As an additional comparison, we also included a robust contrast from a non-linguistic spatial working memory task (see Experiment 2 (Reading lists of nonwords—that vary in their well-formedness—followed by a memory probe) above).

Validation of the language fROIs (all experiments)

To ensure that the language fROIs behave as expected (i.e. show a reliably greater response to the *sentences* condition compared to the *nonwords* condition), we used an across-runs cross-validation procedure (e.g. Nieto-Castañón and Fedorenko 2012). In this analysis, the first run of the localizer was used to define the fROIs, and the second run to estimate the responses (in percent BOLD signal change, PSC, relative to fixation baseline) to the localizer conditions, ensuring independence (e.g. Kriegeskorte et al. 2009); then the second run was used to define the fROIs, and the first run to estimate the responses; finally, the extracted magnitudes were averaged across the two runs to derive a single response magnitude for each of the localizer conditions. Statistical analyses were performed on these extracted PSC values. Namely, for each of the five left-hemisphere language fROIs identified, we fit a linear mixed-effect (LME) regression model, predicting the level of PSC for sentences relative to nonwords. The model included fixed effects for an intercept and a slope variable encoding the difference between sentences and nonwords on top of the common intercept. This scheme was implemented by coding sentences as a $+0.5$ factor and nonwords as a -0.5 factor. The model additionally included random terms for both the intercept and the slope variable encoding the difference between sentence and nonwords, both grouped by participant:

$$\begin{aligned} \mathit{Effect\ size} \sim\ & 1 + \mathit{diff\_sent\_nonwords} \\ & + (1 + \mathit{diff\_sent\_nonwords} \mid \mathit{participant}) \end{aligned}$$

where 1 denotes the intercept, *diff_sent_nonwords* denotes the difference between sentence and nonwords slope variable, encoded as explained above, and *participant* denotes a unique number per participant.

In this coding scheme, the intercept estimate reflects the average PSC response for the sentence and nonword conditions together and the slope variable estimate reflects the difference between the *sentence* and *nonwords* conditions. Therefore, to test the validity of the language fROIs, we examined the values of the fixed intercept and slope variable estimates. Both of these estimates had to be significantly positive. The results were FDR-corrected for the five ROIs. A similar analysis was performed for the five right-hemisphere fROIs.

Analyses of the critical tasks (all experiments)

To estimate the responses in the language fROIs to the conditions of the critical tasks, in each experiment the data from all the runs of the language localizer were used to define the fROIs, and the responses to each condition were then estimated in these regions (in percent BOLD signal change, PSC, relative to fixation baseline). The critical conditions were as follows (see Design, materials, and procedure above): (i) in experiment 1a: visual nonwords from the language localizer, (ii) in experiment 1b: auditory words/nonwords, (iii) in experiment 1c: auditory nonwords, (iv) in experiment 2: five word/nonword conditions parametrically varying in well-formedness, and (v) in experiment 3: two nonword conditions varying in phonological neighborhood size.

For each experiment, we used LME regression models (using Matlab *fitlme* routine) to determine the significance of activations of the critical conditions within the language network. We used these models in two ways: (i) to examine the response within the language network as a whole and (ii) to examine the responses in each of the five language fROIs separately. Treating the language network as an integrated system is reasonable given that the regions of this network (a) show similar functional profiles, both with respect to selectivity for language over non-linguistic processes (e.g. Fedorenko et al. 2011) and with respect to their role in lexico-semantic and syntactic processing (e.g. Blank et al. 2016; Fedorenko et al. 2020), and (b) exhibit strong inter-region correlations in both their activity during naturalistic cognition paradigms (e.g. Blank et al. 2014; Paunov et al. 2018; Braga et al. 2020; Malik-Moraleda et al. 2022) and key functional markers, like the strength of response or the extent of activation in response to language stimuli (e.g. Mahowald and Fedorenko 2016; Mineroff et al. 2018; Lipkin et al. 2022). However, because we wanted to allow for the possibility that language regions might differ in their response to nonwords, as well as in order to examine the robustness of the effects across the language fROIs, we supplement the network-wise analyses with the analyses of the five language fROIs separately.

For each of the five language fROIs, we fit a linear mixed-effect regression model, predicting the level of PSC in the target language fROI in the contrasted conditions.

In the case of modeling a condition with a single level, as in experiments 1a, b and c, which all contained a single critical condition (nonwords), this condition was modeled as the intercept of the model. The intercept estimates are reported as representing the condition. The model then included a fixed effect for the intercept, and a random intercept grouped by participant.

For the network-level analysis, we included a random intercept grouped by fROI:

$$\mathit{Effect\ size} \sim 1 + (1 \mid \mathit{participant}) + (1 \mid \mathit{fROI})$$

For the ROI-level analysis, we ran this model for each ROI:

$$\mathit{Effect\ size} \sim 1 + (1 \mid \mathit{participant})$$

The P-values (comparing the intercept estimate to 0) were FDR-corrected for the 5 ROIs.

In the case of modeling a condition with multiple levels, we added a slope variable encoding the effect of the critical condition beyond the common intercept. For experiment 2, we modeled the five levels of well-formedness by coding them on a linear scale from -2 to 2 (multiplying brain activity by the factors -1, -0.5, 0, 0.5, 1) from low to high well-formedness, respectively. In experiment 3, we coded the low phonological neighborhood condition as -0.5 and the high neighborhood condition as 0.5.

In these cases, the model included fixed effects for the intercept and condition (the slope variable coding the critical condition) and potentially correlated random intercepts and slopes grouped by participant. Here, the intercept represents the mean brain activity across all the levels of the critical condition and the condition slope estimate represents the deviation in brain activity due to the different levels of the critical conditions. Therefore, the overall effect of the critical condition was significant if the condition estimates were significantly different from 0.

For the network-level analysis for experiments 2 and 3 we included potentially correlated random intercept and slopes grouped by fROI:

$$\begin{aligned}\textit{Effect size} \sim 1 + \textit{condition} + (1 + \textit{condition} \mid \textit{participant})\\ + (1 + \textit{condition} \mid \textit{fROI})\end{aligned}$$

For the ROI-level analysis, we ran this model for each fROI:

$$\textit{Effect size} \sim 1 + \textit{condition} + (1 + \textit{condition} \mid \textit{participant})$$

The *P*-values comparing the condition estimates to 0 were FDR-corrected for the five fROIs.

Parallel analyses were run for the right-hemisphere language fROIs.

A similar procedure was applied to evaluate the effects for the fROIs that were selected based on the word/nonword well-formedness gradient in experiment 2.

Results

Validation of the language fROIs (all experiments)

Across all experiments, each of the five left-hemisphere fROIs (see Fig. 2B for parcel locations and names) showed a reliably above-baseline response to *Sentences* (all intercept estimates > 0, $Ps < 0.001$, FDR-corrected for the five fROIs; full results available at OSF: https://osf.io/6c2y7/), as well as a robust *Sentences* > *Nonwords* effect (all slope estimates > 0, $ps < 0.001$, FDR-corrected for the five fROIs), consistent with much previous work (e.g. Fedorenko et al. 2010; Mahowald and Fedorenko 2016; Diachek et al. 2020; Lipkin et al. 2022).

Behavioral measures in the fMRI tasks

The behavioral tasks that we included in some of the fMRI paradigms were designed to maintain participants' alertness throughout the experiment while providing us with quantitative estimates of alertness in the different conditions. In general, the behavioral performance in all the tasks was relatively high (>70%) and revealed no significant differences between experimental conditions. The only exception was experiment 2, where nonword well-formedness had a small effect on the memory probe task performance, with better performance for more well-formed conditions [fixed slope of accuracy as a function of well-formedness in a LME model that includes a fixed effect of condition and random intercepts and slopes for participants; −0.02, $t(78) = -2.23$, $P = 0.025$, Fig. 3C]. See Supplementary Information Section 2 for average behavioral performance for all experiments. Raw behavioral data are available at OSF: https://osf.io/6c2y7/.

Key result 1: The language fROIs respond robustly to visually and auditorily presented nonwords (experiments 1a to c)

In experiment 1a, visually presented nonwords elicited a robust response relative to the fixation baseline across the language network as a whole, when treating the fROIs as a random effect ($P < 0.001$; Table S2, Fig. 2A), and in each of the five fROIs individually ($ps < 0.001$, FDR-corrected; Table S1, Fig. 2B). Similarly, auditorily presented nonwords elicited a robust response relative to the fixation baseline. This result held for the network as a whole in both experiment 1b and experiment 1c ($Ps < 0.01$; Table S2, Fig. 2A). In experiment 1b, this effect was also reliable in each of the five language fROIs ($Ps < 0.05$, FDR-corrected, Table S1, Fig. 2B), and in experiment 1c, this effect was reliable in the two temporal fROIs (AntTemp and PostTemp fROIs; $Ps < 0.01$, FDR-corrected; Table S1, Fig. 2B). Thus, experiments 1a-1c revealed robust sensitivity in the language fROIs to nonwords across modalities and tasks.

It is worth noting that a language-responsive region in the left angular gyrus was not sensitive to nonwords (Fig. S2). The left AngG language fROI was originally included as part of the language network (Fedorenko et al. 2010) but subsequently excluded given its functional differentiation from the rest of the language fROIs (e.g. Shain, Paunov, Chen et al. 2023; see Supplementary Information Section 4 for details). The lack of this fROI's sensitivity to nonwords provides yet another piece of evidence for its distinctness from the core frontal and temporal language regions.

Key result 2: The language fROIs respond more strongly to more well-formed nonwords (experiment 2)

The well-formedness manipulation resulted in a gradient of fMRI response strength in the language network such that words and more well-formed nonwords elicited stronger responses than less well-formed nonwords ($P < 0.001$, Table S1, Fig. 2). This result held both for the network as a whole and in each of the five fROIs individually (all $Ps < 0.001$, FDR-corrected, Tables S1 and S2). Thus, experiment 2 suggested that the language network is strongly sensitive to the well-formedness of nonwords.

To test whether this effect was restricted to the language network, we performed a whole-brain search for regions that show a reliable gradient response to nonword well-formedness (along with an above-baseline response to the most well-formed condition). This search revealed five brain regions: four in the left hemisphere (regions A to D, Fig. 3) and one in the right hemisphere (region E, Fig. 3). The masks for the left-hemisphere regions roughly coincided with the language masks (regions A to D roughly coincided with the PostTemp, IFG, AntTemp, and IFGorb language masks, respectively, see Figs 2 and 3). The right-hemisphere region was very small (only four voxels) and was buried inside the superior temporal sulcus, roughly coinciding with the RH AntTemp language mask (Fig. 3 and Supplementary Information Section 3, Fig. S1). Importantly, all of these regions showed robust sensitivity to language processing: the responses to the *Sentences* condition from the language localizer were significantly larger than to the *Nonwords* condition (all $Ps < 0.0001$, FDR-corrected for the five regions, full results at OSF: https://osf.io/6c2y7/). This result suggests that the brain regions that are most sensitive to the degree of nonword well-formedness across the whole brain are also sensitive to lexical semantics and syntactic/combinatorial processing, and that by focusing on the language-responsive areas in our main analysis, we did not miss any critical areas outside of the language network.

To evaluate the similarity of the fine-grained activation patterns between the nonword well-formedness contrast and the language localizer contrast, we performed two additional analyses. First, we visually examined individual whole-brain activation maps for the two contrasts, along with an additional control contrast from the spatial working memory task (Supplementary Information Section 6, Fig. S4). In line with the results of the whole-brain search for areas sensitive to nonword

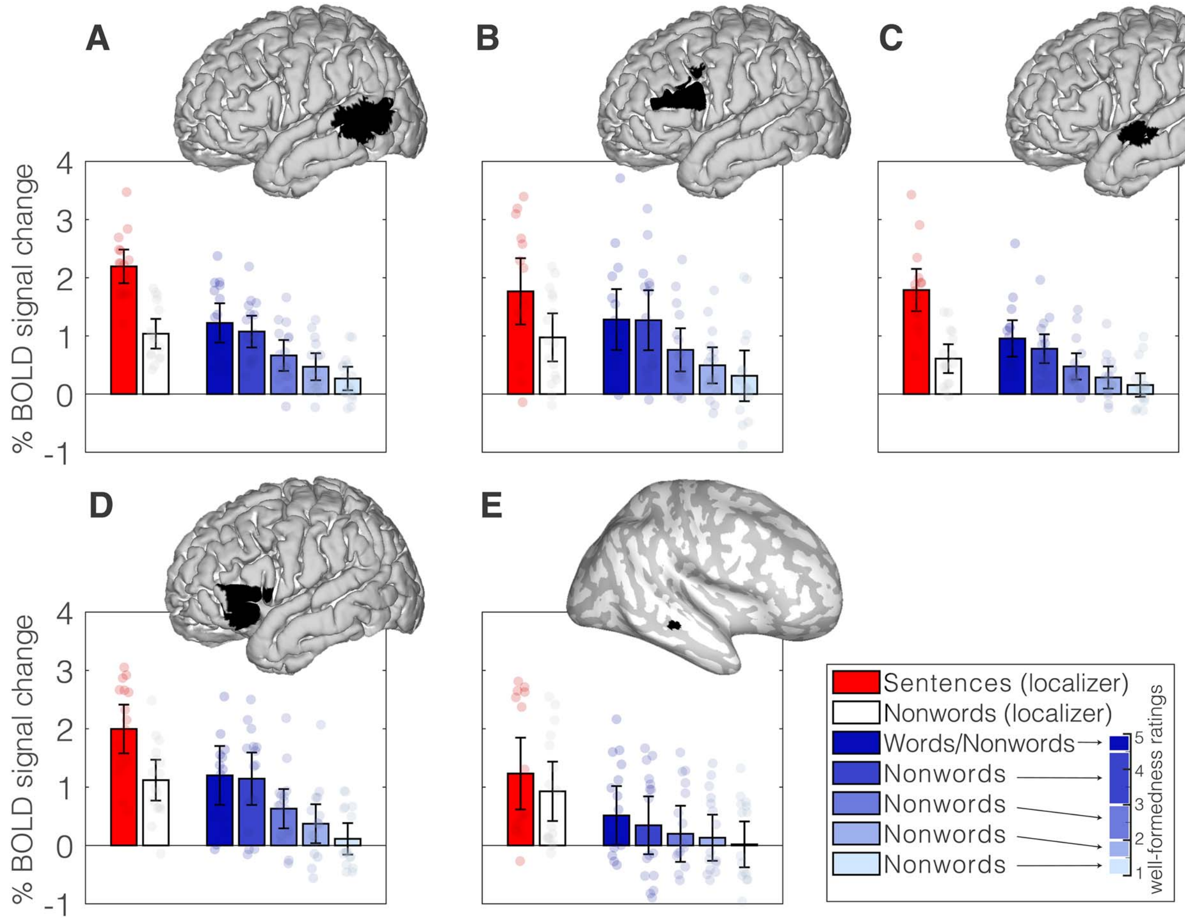

Fig. 3. Brain regions that are sensitive to nonword well-formedness across the brain (experiment 2, $n = 16$). A–E) Five brain regions that were found using a GcSS approach (Fedorenko et al. 2010). (A–D) are left-hemisphere regions and (E) is a right-hemisphere region. (E) is a small parcel (only four voxels) that was buried inside the superior temporal sulcus and is visible only when plotted on top of an unfolded cortical surface. The brain images display the masks that were used to define the fROIs; individual fROIs are 10% of voxels that are most sensitive to the nonword well-formedness gradient within each mask. Bar graphs show % BOLD signal change relative to a fixation baseline in individually defined fROIs averaged across participants; the effects are estimated using data that are independent from the data used to define the fROIs. Bar graphs, all panels, left to right—sentences (red) and nonwords (white) from the language localizer (similar to Experiment 1a) that was run on these 16 participants, 5 conditions from experiment 2, from high to low well-formedness (shades of blue).

well-formedness above, the activations for the nonword well-formedness contrast appear similar to the activations for the language localizer (although the latter is, of course, a broader and more robust contrast, leading to overall stronger activations); in contrast, the control, nonlinguistic spatial working memory task elicits a very different pattern of activations, in line with past work (e.g. Fedorenko et al. 2013a). To quantify this similarity, we computed voxel-wise spatial correlations (Supplementary Information Section 7, Figs. S5 and S6). This analysis asks whether, e.g. the most language-responsive voxels also show the strongest sensitivity to nonword well-formedness. We found a strong correlation between the nonword well-formedness contrast and the language localizer contrast (>0.5, on average across participants, across the whole left hemisphere); in contrast, the correlations between each of the two language contrasts and the spatial working memory contrast are close to zero or negative (Fig. S5). Further, the correlation between the nonword well-formedness contrast and the language localizer contrast was approximately as high as the correlation across the runs of the nonword reading task, representing the noise ceiling (Fig. S6). These results thus strengthen our claim that phonotactic regularities are primarily processed within the language network.

Key result 3: No evidence for lexical "neighbors" driving the language network's response to nonwords (experiments 2 and 3)

One possible explanation for the results of experiment 2 is that reading nonwords that are well-formed activates the representations of real words that are similar to them (e.g. BRIVERY ➔ BRAVERY). Thus, given the strong sensitivity of the high-level language network to word meanings (e.g. Fedorenko et al. 2012b; Pereira et al. 2018), stronger responses to more well-formed nonwords could be explained on a purely lexical basis, without invoking sublexical/phonological regularities.

We tested this possibility in two ways by focusing on phonological neighborhood measures because nonwords that are similar to (and may therefore activate) real words will have higher neighborhood density: *first*, in experiment 3, we measured neural responses to two groups of nonwords that were matched on phonotactic probability [two-sample *t*-test, orthography-based measure: $t(172) = 1.1$, $P = 0.27$, Fig. 4D; see Methods for a pronunciation-based measure yielding similar results] but differed in the size of their orthographic and phonological neighborhood [$t(172) = 6.9$ $P < 0.001$, Fig. 4E], and *second*, we computed the average orthographic neighborhood size of the

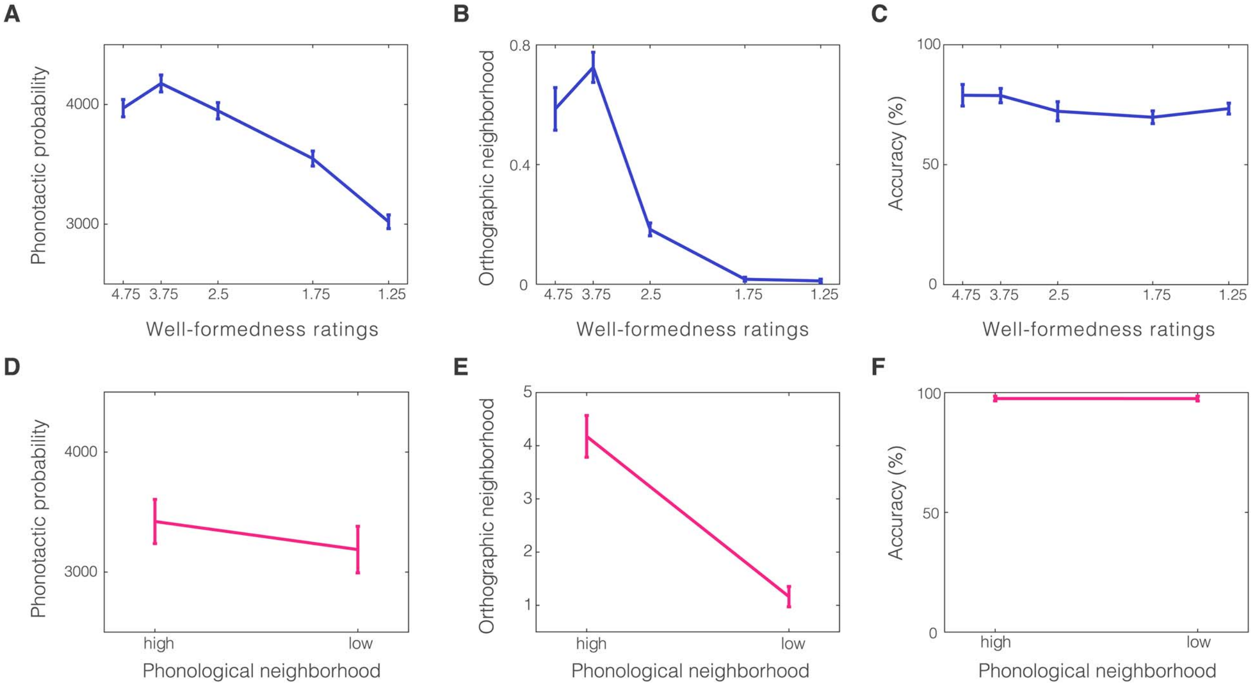

Fig. 4. Stimulus characteristics and behavioral results, Experiments 2 A–C) and 3 D–F). A) Phonotactic probability of the materials in Experiment 2. The ordinate represents phonotactic probability (Methods)—The mean count from an English corpus of all bigrams that occur in a nonword; the abscissa represents the five conditions in Experiment 2, ordered by the bin centers of the well-formedness ratings, from most to least well-formed. Note that the highest well-formed group (bin center 4.75) mostly contained real words, but all four other groups contained only nonwords. B) Orthographic neighborhood size of the materials in Experiment 2. The ordinate represents orthographic neighborhood size (Methods), i.e. the number of real words that are identical to the nonword up to a substitution of a single letter. The abscissa is the same as in (A). C) Behavioral results in experiment 2. The ordinate is the accuracy in the memory probe task (Methods). The abscissa is the same as in (A). D) Phonotactic probability of the materials in Experiment 3. The ordinate is the same as in (A). The abscissa represents the two conditions in Experiment 3. The graph shows a numerical decrease of phonotactic probability due to neighborhood size but this effect is not significant (see text). E) Orthographic neighborhood size of the materials in Experiment 3. The ordinate is the same as in (B). The abscissa is the same as in (D). F) Behavioral results in Experiment 3 (shows participants were at ceiling for both conditions). The ordinate is the accuracy in the repetition detection task (Methods). The abscissa is the same as in (D).

nonwords in the five conditions of experiment 2 and examined the relationship between this measure and neural response strength.

In experiment 3, the high- and low-neighborhood conditions elicited responses that were comparable in magnitude across the language network (pink bars in Fig. 2, experiment 3), with no evidence for stronger responses to high-neighborhood nonwords either in the language network as a whole or in any of the individual fROIs (*ps* > 0.1, Table S1 and 3).

In experiment 2, of greatest relevance are the two conditions with the lowest well-formedness ratings (the two rightmost, light blue bars in Fig. 2, experiment 2). Although these conditions have similarly low orthographic neighborhood size [both around 0; two-sample *t*-test: $t(718) = 0.6$, $P = 0.52$, Fig. 4B], they elicited differential brain responses such that the second-lowest well-formed condition activated the language network significantly more than the least well-formed condition [a post-hoc LME revealed a small but significant difference in PSC between these two conditions = 0.14, $t(158) = 2.4$, $P = 0.017$]. In contrast to their similar orthographic neighborhood size, these conditions differ reliably in their phonotactic probability [$t(718) = 6.2$, $P < 0.001$; Fig. 4A], largely mirroring the well-formedness ratings.

In summary, the results of both experiment 3 and the post-hoc analysis of the two least well-formed conditions in experiment 2 suggest that phonotactic probability (likely reflected in the well-formedness ratings in experiment 2) explains neural responses in the language regions better than neighborhood size. This result suggests that these responses are not likely to be due to the activation of lexical representations of neighboring real words.

Discussion

Across five fMRI experiments, we investigated the responses of "high-level" language processing brain regions (Fedorenko et al. in press) to nonwords—meaningless sequences of sounds/letters (e.g. punes, silory, flope)—and found that these regions indeed robustly respond to such stimuli in an abstract (modality- and task-independent) fashion. Moreover, we found that the language regions are highly sensitive to the phototactic well-formedness of nonwords, which suggests that regions that extract high-level meaning from language also represent and process sublexical phoneme-combinatorial regularities. In the remainder of the discussion, we situate these findings in the broader theoretical and empirical context and discuss their implications.

The high-level language network responds to nonwords

A network of frontal and temporal brain regions supports language processing. These regions respond during both listening to and reading of linguistic stimuli (e.g. Fedorenko et al. 2010; Vagharchakian et al. 2012; Regev et al. 2013; Scott et al. 2017) across tasks (e.g. Fedorenko et al. 2010; Cheung et al. 2020; Diachek et al. 2020), but show little or no response to diverse non-linguistic functions (Fedorenko et al. 2011; Monti et al. 2012; Ivanova et al. 2020, 2021; Fedorenko and Blank 2020).

The precise contributions of this network to language processing remain debated (Hickok and Poeppel 2007; Price 2010; Friederici 2011; Indefrey 2011; Hagoort 2013, 2019; Duffau et al.

2014; Pylkkänen 2019; Fedorenko et al. in press). Many have argued that distinct subsets of this network store and process syntactic/-combinatorial structure vs. word meanings (e.g. Grodzinsky and Santi 2008; Baggio and Hagoort 2011; Friederici 2011, 2012; Tyler et al. 2011; Duffau et al. 2014; Ullman 2015). However, evidence has been accumulating against this distinction, suggesting that each region of the language network supports both syntactic and lexico-semantic processing (e.g. Dick et al. 2001; Wilson and Saygin 2004; Fedorenko et al. 2010, 2012, 2020; Bautista and Wilson 2016; Blank et al. 2016; Shain, Kean et al. 2023). Other work has implicated the language network in word-internal morphological processing (e.g. Bozic et al. 2010).

The current study establishes that the language network is sensitive to an even shorter scale of linguistic information relative to syntax, lexical semantics, and morphology—sublexical sound patterns—as evidenced by responses to sequences of phonemes that do not constitute real words. The response to nonwords in the language network is, by definition, lower than the response to sentences because this network is defined by the *Sentences* > *Nonwords* contrast (Fedorenko et al. 2010). Nevertheless, nonwords elicit a response that is consistently and reliably higher than the low-level baseline. Above-baseline responses to nonwords in the language network can be observed in prior fMRI (e.g. Fedorenko et al. 2010; Mahowald and Fedorenko 2016; Mollica et al. 2020; Chen et al. 2023) and intracranial (e.g. Fedorenko et al. 2016) reports. Additionally, previous data show that the responses to nonwords are larger than to many nonlinguistic tasks, including arithmetic, spatial working memory, and music perception (e.g. Mineroff et al. 2018; Fedorenko and Blank 2020; Chen et al. 2023). Sensitivity of the language network to phonological information is also consistent with reliable responses to unfamiliar foreign languages—from which only phonological-level information can be extracted—in the language regions of bilinguals and polyglots (Malik-Moraleda et al. 2022, Malik-Moraleda, Jouravlev et al. 2023) and with robust representations of phonemic information across the language network during naturalistic auditory language comprehension (e.g. Gong et al. 2023). However, this is the first study to systematically investigate the responses in the language network to nonwords and try to understand what drives them.

The fact that the language regions respond both when participants read nonwords (experiments 1a, 2, and 3) and when they listen to them (experiments 1b and 1c) demonstrates that the representation of nonwords is abstract (unpublished findings from Rebecca Saxe's lab further show that nonwords in American Sign Language (ASL)—signs similar in form to meaningful ones but lacking meaning—also elicit above-baseline responses in the language areas; data as published in Richardson et al. 2020). These results align with previous findings of modality-independent responses of the language network to stories, sentences, and word lists (e.g. Fedorenko et al. 2010; Vagharchakian et al. 2012; Regev et al. 2013), but critically extend them to stimuli that lack meaning.

Similarly, we show that the response to nonwords in the language network generalizes across tasks, including passive reading/listening, processing of nonword strings followed by a memory probe ("did you encounter this nonword in the preceding string?"), and repetition detection. These findings are in line with the task-independence of the language network's responses to words and sentences (e.g. Cheung et al. 2020; Diachek et al. 2020). Importantly, none of these tasks require selective attention to particular properties of the nonwords, which suggests that this response reflects the *intrinsic computations* necessary for recognizing and processing sublexical sound patterns. Such computations are presumably critical to language acquisition and processing given that any newly encountered word is, at first, just a sequence of sounds that gradually acquires semantic associations as we learn the word's meaning (e.g. Davis et al. 2009; Perry et al. 2018; Jones et al. 2021).

Combined with prior studies, our results suggest that the fronto-temporal language network supports not only the processing of words and inter-word dependencies but also of lower-level phonological information, as evidenced by strong responses to sequences of phonemes that obey phoneme-combinatorial constraints but do not correspond to meanings in our lexicon. The reports of phonological impairments following brain lesions that also cause higher-level linguistic deficits, i.e. aphasia (e.g. Geva et al. 2011; Kries et al. 2023) further point to a *causal role* of these brain areas in phonological processing. Any proposal about the language network's computations should therefore account for its role in phonological-level processing.

The language network is sensitive to phonotactic regularities

In experiment 2, we found that more well-formed/phonotactically probable nonwords elicit stronger responses in the language network. This modulation of neural activity by nonword well-formedness plausibly reflects a process of matching sound patterns to stored representations extracted from our previous experience with a language, whereby the strength of response is proportional to how well the stimulus matches stored linguistic regularities and the amount of matching information (e.g. Hayes and Wilson 2008). This idea is reminiscent of the notion of "phonological schemata" (Jackendoff 2002). Storage of frequent sound/letter n-grams may allow for more efficient processing through enabling the representation assembly to proceed in larger chunks than single phonemes/letters (Bybee 1999; Bybee and Hopper 2001; Vitevitch and Luce 2005; O'Donnell 2015).

Might the stronger responses to more well-formed nonwords instead (or additionally) reflect activation of lexical representations of real words that share phonological/sound structure with them? We evaluated this possibility in two ways and did not find support for it. First, in experiment 3, nonwords that differed in the number of real-word neighbors (but were matched on phonotactic probability) elicited similar-magnitude responses in the language network. Second, in experiment 2, we found that although the two least well-formed groups of nonwords both had few or no real-word neighbors, the more well-formed nonwords elicited stronger responses in the language network. So, it appears that the language regions represent sublexical units including phoneme sequences that are not associated with a lexical–semantic representation. We suggest that the frequency of these phoneme sequences in our experience with the language is what drives the response to nonwords in the language regions, even if familiar sound patterns do not lead to lexical activation of similar-sounding real words.

A possibility that is more difficult to rule out is that the response to nonwords is, at least in part, driven by the (relatively rare) semantic associations that might be elicited by particular sounds/sound clusters (e.g. Iwasaki et al. 2007; Monaghan et al. 2014; Larsson 2015; Blasi et al. 2016; Winter et al. 2017; Sidhu and Pexman 2018; Pimentel et al. 2019; Vinson et al. 2021) or morphemes/morpheme-like elements that occur in some nonwords (e.g. Bozic et al. 2010). Further research is needed to determine the precise features that make a nonword elicit an above-baseline response in high-level language areas, including whether sublexical semantic associations may be sufficient to explain this response. In addition, developmental investigations, especially during the first few years of life—when most words we encounter do not yet have meaning—could help illuminate the formation of linguistic knowledge representations (e.g. Jones et al. 2021).

Phonological processing outside of the high-level language network?

In a whole-brain search, all the brain regions that showed sensitivity to nonword well-formedness also showed sensitivity to high-level linguistic meaning, suggesting that they fall within the boundaries of the language network. Furthermore, whole-brain voxel-wise activation patterns were highly similar between the nonword well-formedness contrast and the language localizer contrast. These results suggest that not only are language regions *sensitive* to phonotactic regularities, they also constitute the *primary processing system* for these regularities. However, it is important to clarify that our core claim is a *positive claim* about the sensitivity of the language regions to sublexical regularities. We are not making a strong argument about the *lack of sensitivity* to phonotactic regularities, or more general contributions to phonological processing, by brain regions outside of the language network. The latter claim would require additional evidence, such as (1) a more comprehensive characterization of the functional profiles of the phonology-sensitive regions we found in the whole-brain search and/or (2) functionally identifying specific brain regions other than the language network in individual participants (some candidate regions are mentioned below) and examining their responses to diverse kinds of nonwords under different task conditions.

Aside from the language regions, where might one expect to find sensitivity to phonotactic well-formedness? One likely candidate are the *speech perception areas* in the superior temporal gyrus. These areas are highly selective for the processing of speech sounds relative to other sounds (e.g., Norman-Haignere et al. 2015), represent the identity of single phonemes (e.g., Mesgarani et al. 2014; Leonard, Gwilliams et al. 2023) and have even been suggested to be sensitive to transitional probabilities in multi-phoneme sequences (Leonard et al. 2015). Importantly, these areas are distinct from the language areas (Fedorenko et al. in press): unlike the language areas, the speech perception areas are not sensitive to linguistic meaning, showing similarly strong response to meaningful and meaningless speech (e.g. Norman-Haignere et al. 2015; Overath et al. 2015). Given that these areas have relatively short "temporal receptive windows" (e.g. Hasson et al. 2008) of approximately half a second (e.g. Overath et al. 2015; Norman-Haignere et al. 2022), they plausibly process temporally local phonological information and provide input to the language areas, which integrate information across longer scales—syllables and words—and compute linguistic meaning (Lerner et al. 2011; Blank and Fedorenko 2020; Regev et al. 2023). It is therefore possible that we did not see sensitivity to phonotactic well-formedness in the speech perception areas because the nonwords in experiment 2 were locally well-formed (by design). Another possibility is that the speech perception areas are only sensitive to phonotactic regularities for auditorily presented sequences (whereas our stimuli were presented visually). The latter would imply that these areas' representations and computations are not truly abstract (in contrast to the language areas) and are instead tied specifically to the auditory modality. A definitive answer would require a version of experiment 2 with both visual and auditory stimuli (ideally, with manipulations that affect well-formedness at different temporal scales) and an independent functional localizer for speech perception areas (e.g. Overath et al. 2015).

In addition to the language areas and speech perception areas, past neuroimaging and patient studies of phonological processing have implicated a wide array of cortical, subcortical, and cerebellar areas (e.g. see Vigneau et al. 2006; Price 2012), including studies that, similar to the current study, have used manipulations of phonotactic/orthographic well-formedness and phonological/orthographic neighborhoods (e.g. Okada and Hickok 2006; Vinckier et al. 2007; Vaden et al. 2011; Gow and Nied 2014; Gow and Olson 2015; Woolnough et al. 2020; Avcu et al. 2023). However, several factors make it challenging to interpret these findings and to relate them to the current results. First, most prior studies have used a single set of stimuli and a single task, making it difficult to assess the robustness and generalizability of the reported results. Second, many of the tasks that are commonly used in investigations of phonological processing go beyond the natural "task" of processing linguistic input with the goal of meaning extraction. As a result, these tasks may engage cognitive processes, and associated neural mechanisms, beyond those that support the processing of linguistic input.

For example, tasks like rhyme judgments (e.g. Petersen et al. 1989; Paulesu et al. 1993; Seghier et al. 2004; Geva et al. 2011; Pillay et al. 2014; Yen et al. 2019), nonword repetition (e.g. Fridriksson et al. 2010; Church et al. 2011; Scott and Perrachione 2019), or other tasks that require active maintenance of words/nonwords in working memory (e.g. Paulesu et al. 1993; Awh et al. 1996) may engage the *articulation network* (e.g. Bohland and Guenther 2006; Guenther 2016; Basilakos et al. 2017, 2018). Similar to the speech perception areas—and in contrast to the language areas—the articulation areas are only sensitive to the surface properties of speech, not to linguistic meaning (Fedorenko et al. in press). Some phonological tasks may instead, or in addition, engage areas of the domain-general *MD network*, which supports task demands across domains (e.g. Duncan 2010, 2013; Fedorenko et al. 2013a; Shashidhara et al. 2019), is robustly distinct from the language network (e.g. Fedorenko et al. 2012a; Fedorenko and Blank 2020) and has been shown to get engaged when linguistic processing is accompanied by extraneous task demands (Diachek et al. 2020; Quillen et al. 2021). Importantly, however, because past work has not relied on functional localizers, interpretation of activation in a particular anatomical area as reflecting a particular perceptual, motor, or cognitive process—what is known as a "reverse inference"—is challenging (Poldrack 2006; Fedorenko 2021).

Conclusion

We have presented evidence that auditory or visual meaningless sequences of phonemes elicit responses in the language network—a set of brain regions that have been traditionally associated with the processing of word meanings and word-combinatorial processing. This robust sensitivity of the high-level language regions to sublexical phonemic patterns aligns with views of linguistic knowledge and processing where the boundaries between different levels of linguistic structure—from phonemes to morphemes to words to constructions and syntactic rules—are not sharp (e.g. Gaskell and Marslen-Wilson 1997; Bybee 1999, 2013; Goldberg 2003; Jackendoff 2007; Huettig et al. 2020; Jackendoff and Audring 2020), and it challenges accounts of the language network, or its subcomponents, that focus on phrase-structure building, compositional meaning, or prediction at the level of word sequences.

Acknowledgments

We would like to acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, and its support team (Steve Shannon and Atsushi Takahashi). We thank former and current EvLab members for their help with fMRI data collection, Steve Piantadosi for help in creating the materials for experiment 2, Peter Graff for early discussions of phonology in the brain, Alex Paunov for help with some of the analyses, and Ray Jackendoff for helpful comments on this work. T.I.R. also thanks Janet Werker, the audience at the 2020 Neurobiology of Language conference, and Tali Bitan and her lab for helpful discussions.

Author contributions

Conceptualization: T.I.R., L.B., K.M., E.F. Data curation: T.I.R., A.E.S, L.B., K.M., E.F. Investigation: J.A., K.M., E.F. Formal analysis: T.I.R., H.S.K., X.C., J.A., E.F. Visualization: T.I.R Writing-original draft: T.I.R., K.M., E.F. Writing - review & editing: all authors. Supervision: K.M., E.F.

Supplementary material

Supplementary material is available at *Cerebral Cortex* online.

Funding

T.I.R. was supported by the Zuckerman-CHE STEM Leadership Program and the Poitras Center for Psychiatric Disorders Research. E.F. was supported by the National Institute of Health (NIH): R00 award HD057522, R01 awards DC016607 and DC016950, and funds from the Brain and Cognitive Sciences department, the McGovern Institute for Brain Research, and the Simons Center for the Social Brain.

Conflict of interest statement: None declared.

Data availability

Processed data and materials are available at: Open Science Framework (OSF) platform (https://osf.io/6c2y7/). Raw data can be made available upon request.

Abbreviations

AntTemp—anterior temporal; PostTemp—posterior temporal; IFG—inferior frontal gyrus; IFGorb—inferior frontal gyrus, orbital portion; MFG—medial frontal gyrus; AngG—angular gyrus; AntTemp-L—anterior temporal, left hemisphere (and similarly for PostTemp-L, IFG-L, IFGorb-L, MFG-L and AngG-L); AntTemp-R—anterior temporal, right hemisphere (and similarly for PostTemp-R, IFG-R, IFGorb-R, MFG-R and AngG-R); GcSS analysis—group-constrained subject-specific analysis; fROI—functional region of interest; fMRI—functional magnetic resonance imaging.

References

Albright A. How many grammars am I holding up? Discovering phonological differences between word classes; In *Proceedings of the 26th west coast conference on formal linguistics* 2008 (pp. 1–20).

Arciuli J, Monaghan P. Probabilistic cues to grammatical category in English orthography and their influence during reading. *Scientific Studies of Reading*. 2009:13:73–93. http://dx.doi.org/101080/10888430802633508.

Arciuli J, McMahon K, de Zubicaray G. Probabilistic orthographic cues to grammatical category in the brain. *Brain Lang*. 2012:123(3): 202–210.

Ashburner J, Friston KJ. Unified segmentation. *NeuroImage*. 2005:26(3):839–851.

Avcu E, Newman O, Ahlfors SP, Gow DW. Neural evidence suggests phonological acceptability judgments reflect similarity, not constraint evaluation. *Cognition*. 2023:230:105322.

Awh E, Jonides J, Smith EE, Schumacher EH, Koeppe RA, Katz S. Dissociation of storage and rehearsal in verbal working memory: evidence from positron emission tomography. *Psychol Sci*. 1996:7(1):25–31.

Baggio G, Hagoort P. The balance between memory and unification in semantics: a dynamic account of the N400. *Lang Cogn Process*. 2011:26(9):1338–1367.

Balota DA, Yap MJ, Cortese MJ, Hutchison KA, Kessler B, Loftis B, Neely JH, Nelson DL, Simpson GB, Treiman R. The english lexicon project. *Behav Res Methods*. 2007:39(3):445–459.

Basilakos A, Fridriksson J, Rorden C, Behroozmand R, Hanayik T, Naselaris T, Del GJ, Breedlove J, Vandergrift WA, Bonilha L. Activity associated with speech articulation measured through direct cortical recordings. *Brain Lang*. 2017:169:1–7.

Basilakos A, Smith KG, Fillmore P, Fridriksson J, Fedorenko E. Functional characterization of the human speech articulation network. *Cereb Cortex*. 2018:28(5):1816–1830.

Bautista A, Wilson SM. Neural responses to grammatically and lexically degraded speech. *Lang Cogn Neurosci*. 2016:31(4):567–574.

Berwick RC, Chomsky N. Why only us: Language and evolution. MIT press; 2016.

Blank IA, Fedorenko E. No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*. 2020:219:116925.

Blank I, Kanwisher N, Fedorenko E. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *J Neurophysiol*. 2014:112(5): 1105–1118.

Blank I, Balewski Z, Mahowald K, Fedorenko E. Syntactic processing is distributed across the language system. *NeuroImage*. 2016:127: 307–323.

Blasi DE, Wichmann S, Hammarström H, Stadler PF, Christiansen MH. Sound-meaning association biases evidenced across thousands of languages. *Proc Natl Acad Sci USA*. 2016:113(39): 10818–10823.

Boatman D. Cortical bases of speech perception: evidence from functional lesion studies. *Cognition*. 2004:92(1–2):47–65.

Bohland JW, Guenther FH. An fMRI investigation of syllable sequence production. *NeuroImage*. 2006:32(2):821–841.

Bozic M, Tyler LK, Ives DT, Randall B, Marslen-Wilson WD. Bihemispheric foundations for human speech comprehension. *Proc Natl Acad Sci USA*. 2010:107(40):17439–17444.

Bromberger S, Halle M. Why phonology is different. *Linguistic inquiry*. 1989;20(1):51–70.

Braga RM, DiNicola LM, Becker HC, Buckner RL. Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *J Neurophysiol*. 2020:124(5):1415–1448.

Burton MW. The role of inferior frontal cortex in phonological processing. *Cogn Sci*. 2001:25(5):695–709.

Bybee J. Usage-based phonology. In: *Functionalism and formalism in linguistics: volume I: general papers*. John Benjamins Publishing; 1999. 1:211–242. https://www.torrossa.com/en/resources/an/5001357#page=218.

Bybee J. Usage-based theory and exemplar representations of constructions. In: Hoffmann T, Trousdale G (eds), *The Oxford* *handbook of construction grammar*; Oxford Academic, 2013. pp. 49–69. (accessed 23 Feb. 2024). https://doi.org/10.1093/oxfordhb/9780195396683.013.0004.

Bybee JL, Hopper PJ, editors. *Frequency and the emergence of linguistic structure (review)*. John Benjamins Publishing Company; 2001; 1–502. https://www.torrossa.com/en/resources/an/5002168.

de Carvalho A, Dautriche I, Christophe A. Preschoolers use phrasal prosody online to constrain syntactic analysis. *Dev Sci*. 2016:19(2): 235–250.

Celsis P, Boulanouar K, Doyon B, Ranjeva JP, Berry I, Nespoulous JL, Chollet F. Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. *NeuroImage*. 1999:9(1):135–144.

Chen X, Affourtit J, Ryskin R, Regev TI, Norman-Haignere S, Jouravlev O, Malik-Moraleda S, Kean H, Varley R, Fedorenko E. The human language system, including its inferior frontal component in “Broca’s area,” does not support music perception. *Cerebral Cortex*. 2023:bhad087.

Chen X, Affourtit J, Ryskin R, Regev TI, Norman-Haignere S, Jouravlev O, Malik-Moraleda S, Kean H, Varley R, Fedorenko E. The human language system, including its inferior frontal component in “Broca’s area,” does not support music perception. *Cerebral Cortex*. 2023;bhad087.

Cheung C, Ivanova A, Siegelman M, Pongos A, Kean H, Fedorenko E. The effect of task on sentence processing in the brain. *Poster presentation at the Society for the Neurobiology of language*; 2020.

Chomsky N. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press; 1965.

Chomsky N. Language and nature. *Mind*. 1995:104(413):1–61.

Chomsky N, Halle M. Some controversial questions in phonological theory. *J Linguist*. 1965:1(2):97–138.

Church JA, Balota DA, Petersen SE, Schlaggar BL. Manipulation of length and lexicality localizes the functional neuroanatomy of phonological processing in adult readers. *J Cogn Neurosci*. 2011:23(6):1475–1493.

Coady JA, Aslin RN. Young children’s sensitivity to probabilistic phonotactics in the developing lexicon. *J Exp Child Psychol*. 2004:89(3):183–213.

Coltheart M, editor. Attention and performance XII: The psychology of reading. *Routledge*; 2016.

Dautriche I, Swingley D, Christophe A. Learning novel phonological neighbors: syntactic category matters. *Cognition*. 2015:143:77–86.

Dautriche I, Mahowald K, Gibson E, Christophe A, Piantadosi ST. Words cluster phonetically beyond phonotactic regularities. *Cognition*. 2017:163:128–145.

Davis MH, Di Betta AM, Macdonald MJE, Gaskell MG. Learning and consolidation of novel spoken words. *J Cogn Neurosci*. 2009:21(4): 803–820.

Demonet J-F, Price C, Wise R, Frackowiak RSJ. A PET study of cognitive strategies in normal subjects during language tasks influence of phonetic ambiguity and sequence processing on phoneme monitoring. *Brain* 1994;117(4):671–682.

Devlin JT, Matthews PM, Rushworth MFS. Semantic processing in the left inferior prefrontal cortex: a combined functional magnetic resonance imaging and transcranial magnetic stimulation study. *J Cogn Neurosci*. 2003:15(1):71–84.

DeWitt I, Rauschecker JP. Phoneme and word recognition in the auditory ventral stream. *Proc Natl Acad Sci USA*. 2012:109(8):2709.

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

Dick F, Bates E, Utman JA, Wulfeck B, Dronkers N, Gernsbacher MA. Language deficits, localization, and grammar: evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychol Rev*. 2001:108(4): 759–788.

Duffau H, Moritz-Gasser S, Mandonnet E. A re-examination of neural basis of language processing: proposal of a dynamic hodotopical model from data provided by brain stimulation mapping during picture naming. *Brain Lang*. 2014:131:1–10.

Duncan J. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends Cogn Sci*. 2010:14(4):172–179.

Duncan J. The structure of cognition: attentional episodes in mind and brain. *Neuron*. 2013:80(1):35–50.

Fedorenko E. The role of domain-general cognitive control in language comprehension. *Front Psychol*. 2014:5:335.

Fedorenko E. The early origins and the growing popularity of the individual-subject analytic approach in human neuroscience. *Curr Opin Behav Sci*. 2021:40:105–112.

Fedorenko E, Blank IA. Broca’s area is not a natural kind. *Trends Cogn Sci*. 2020:24(4):270–284.

Fedorenko E, Hsieh PJ, Nieto-Castañón A, Whitfield-Gabrieli S, Kanwisher N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J Neurophysiol*. 2010:104(2):1177–1194.

Fedorenko E, Behr MK, Kanwisher N. Functional specificity for high-level linguistic processing in the human brain. *Proc Natl Acad Sci USA*. 2011:108(39):16428–16433.

Fedorenko E, Duncan J, Kanwisher N. Language-selective and domain-general regions lie side by side within Broca’s area. *Curr Biol*. 2012a:22(21):2059–2062.

Fedorenko E, Nieto-Castañón A, Kanwisher N. Syntactic processing in the human brain: what we know, what we don’t know, and a suggestion for how to proceed. *Brain Lang*. 2012b:120(2):187–207.

Fedorenko E, Duncan J, Kanwisher N. Broad domain generality in focal regions of frontal and parietal cortex. *Proc Natl Acad Sci*. 2013a:110(41):16616–16621.

Fedorenko E, Hsieh PJ, Balewski Z. A possible functional localiser for identifying brain regions sensitive to sentence-level prosody. *Lang Cogn Neurosci*. 2013b:30(1–2):120–148.

Fedorenko E, Scott TL, Brunner P, Coon WG, Pritchett B, Schalk G, Kanwisher N. Neural correlate of the construction of sentence meaning. *Proc Natl Acad Sci USA*. 2016:113(41):E6256–E6262.

Fedorenko E, Blank IA, Siegelman M, Mineroff Z. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*. 2020:203:104348.

Fedorenko E, Ivanova AA, Regev TI. The language network as a natural kind within the broader landscape of the human brain. *Nat Rev Neurosci*. in press.

Frauenfelder UH, Baayen RH, Hellwig FM, Schreuder R. Neighborhood density and frequency across languages and modalities. *J Mem Lang*. 1993:32(6):781–804.

Fridriksson J, Kjartansson O, Morgan PS, Hjaltason H, Magnusdottir S, Bonilha L, Rorden C. Impaired speech repetition and left parietal lobe damage. *J Neurosci*. 2010:30(33):11057–11061.

Friederici AD. The brain basis of language processing: from structure to function. *Physiol Rev*. 2011:91(4):1357–1392.

Friederici AD. The cortical language circuit: from auditory perception to sentence comprehension. *Trends Cogn Sci*. 2012:16(5): 262–268.

Friston KJ, Ashburner J, Frith CD, Poline J-B, Heather JD, Frackowiak RSJ. Spatial registration and normalization of images. *Hum Brain Mapp*. 1995:3(3):165–189.

Gaskell MG, Marslen-Wilson WD. Integrating form and meaning: a distributed model of speech perception. *Lang Cogn Process*. 1997:12(5–6):613–656.

Geva S, Jones PS, Crinion JT, Price CJ, Baron JC, Warburton EA. The neural correlates of inner speech defined by voxel-based lesion-symptom mapping. *Brain*. 2011:134(10):3071–3082.

Goldberg AE. Constructions: a new theoretical approach to language. *Trends Cogn Sci*. 2003:7(5):219–224.

Gong XL, Huth AG, Deniz F, Johnson K, Gallant JL, Theunissen FE. (2023)Phonemic segmentation of narrative speech in human cerebral cortex. *Nat Commun*. 2023:141(14):1–17.

Gow DW, Nied AC. Rules from words: a dynamic neural basis for a lawful linguistic process. *PLoS One*. 2014:9(1):e86212.

Gow DW, Olson BB. Lexical mediation of phonotactic frequency effects on spoken word recognition: a granger causality analysis of MRI-constrained MEG/EEG data. *J Mem Lang*. 2015:82: 41–55.

Graves WW, Grabowski TJ, Mehta S, Gordon JK. A neural signature of phonological access: distinguishing the effects of word frequency from familiarity and length in overt picture naming. *J Cogn Neurosci*. 2007:19(4):617–631.

Graves WW, Grabowski TJ, Mehta S, Gupta P. The left posterior superior temporal gyrus participates specifically in accessing lexical phonology. *J Cogn Neurosci*. 2008:20(9):1698–1710.

Grodzinsky Y, Santi A. The battle for Broca's region. *Trends Cogn Sci*. 2008:12(12):474–480.

Guenther FH. Neural control of speech. *Neural Control Speech*. Mit Press, 2016.

Hagoort P. MUC (memory, unification, control) and beyond. *Front Psychol*. 2013:4:416.

Hagoort P. The neurobiology of language beyond single-word processing. *Science*. 2019:80(6461):366–358.

Hartwigsen G, Weigel A, Schuschan P, Siebner HR, Weise D, Classen J, Saur D. Dissociating Parieto-frontal networks for phonological and semantic word decisions: a condition-and-perturb TMS study. *Cereb Cortex*. 2016:26(6):2590–2601.

Hayes B. BLICK : a phonotactic probability calculator (manual). 2012. https://linguistics.ucla.edu/people/hayes/BLICK/BLICKManual.pdf.

Hayes B, Wilson C. A maximum entropy model of phonotactics and phonotactic learning. *Linguist Inq*. 2008:39(3):379–440.

Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*. 2008;28(10):2539–2550.

Heinz J, Idsardi W. Sentence and word complexity. *Science*. 2011; 333(6040):295–297.

Heinz J, Idsardi W. What complexity differences reveal about domains in language. *Topics in cognitive science*. 2013;5(1):111–131.

Hickok G, Poeppel D. The cortical organization of speech processing. *Nat Rev Neurosci*. 2007:8(5):393–403.

Huettig F, Audring J, Jackendoff R. A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition*. 2022;224:105050.

Indefrey P. The spatial and temporal signatures of word production components: a critical update. *Front Psychol*. 2011:2:255.

Ivanova AA, Srikant S, Sueoka Y, Kean HH, Dhamala R, O'Reilly UM, Bers MU, Fedorenko E. Comprehension of computer code relies primarily on domain-general executive brain regions. *elife*. 2020:9: 1–24.

Ivanova AA, Mineroff Z, Zimmerer V, Kanwisher N, Varley R, Fedorenko E. The language network is recruited but not required for nonverbal event semantics. *Neurobiol Lang*. 2021:2(2): 176–201.

Iwasaki N, Vinson DP, Vigliocco G. What do English speakers know about Gera-Gera and yota-yota?: a cross-linguistic investigation of mimetic words of laughing and walking. *Japanese-language Educ around globe*. 2007:17:53–78.

Jackendoff R. *Foundations of language: brain, meaning, grammar, evolution*. Ox: Oxford; 2002.

Jackendoff R. A parallel architecture perspective on language processing. *Brain Res*. 2007:1146:2–22.

Jackendoff R, Audring J. Morphology and memory: toward an integrated theory. *Top Cogn Sci*. 2020:12(1):170–196.

Jones G, Cabiddu F, Andrews M, Rowland C. Chunks of phonological knowledge play a significant role in children's word learning and explain effects of neighborhood size, phonotactic probability, word frequency and word length. *J Mem Lang*. 2021:119:104232.

Kelly MH. Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. *Psychol Rev*. 1992:99(2):349–364.

Keuleers E, Brysbaert M. Wuggy: a multilingual pseudoword generator. *Behav Res Methods*. 2010:42(3):627–633.

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*. 2009:12(5):535–540.

Kries J, De Clercq P, Lemmens R, Francart T, Vandermosten M. Acoustic and phonemic processing are impaired in individuals with aphasia. *Sci Rep*. 2023:13(1):1–15.

Landauer TK, Streeter LA. Structural differences between common and rare words: failure of equivalence assumptions for theories of word recognition. *J Verbal Learning Verbal Behav*. 1973:12(2): 119–131.

Larsson M. Tool-use-associated sound in the evolution of language. *Anim Cogn*. 2015:18(5):993–1005.

Leonard MK, Bouchard KE, Tang C, Chang EF. Dynamic encoding of speech sequence probability in human temporal cortex. *J Neurosci*. 2015:35(18):7203–7214.

Leonard MK, Gwilliams L, Sellers KK, Chung JE, Xu D, Mischler G, Mesgarani N, Welkenhuysen M, Dutta B, Chang EF. Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nat*. 2023:2023:1–10.

Lerner Y, Honey CJ, Silbert LJ, Hasson U. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J Neurosci*. 2011:31(8):2906–2915.

Lipkin B, Tuckute G, Affourtit J, Small H, Mineroff Z, Kean H, Jouravlev O, Rakocevic L, Pritchett B, Siegelman M, et al. Probabilistic atlas for the language network based on precision fMRI data from> 800 individuals. *Scientific Data*. 2022;9(1):529.

Lopopolo A, Frank SL, Van Den Bosch A, Willems RM. Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain Allen P, ed. *PLoS One*. 2017:12(5):e0177794.

Luce PA, Large NR. Phonotactics, density, and entropy in spoken word recognition. *Lang Cogn Process*. 2001:16(5–6):565–581.

Mahowald K, Fedorenko E. Reliable individual-level neural markers of high-level language processing: a necessary precursor for relating neural variability to behavioral and genetic variability. *NeuroImage*. 2016:139:74–93.

Mahowald K, Dautriche I, Gibson E, Piantadosi ST. Word forms are structured for efficient use. *Cogn Sci*. 2018:42(8):3116–3134.

Matchin WG. A neuronal retuning hypothesis of sentence-specificity in Broca's area. *Psychonomic Bulletin & Review*. 2018;25:1682–1694.

Malik-Moraleda S, Jouravlev O, Mineroff Z, Cucu T, Taliaferro M, Mahowald K, Blank IA, Fedorenko E. Functional characterization of the language network of polyglots and hyperpolyglots with precision fMRI. *BioRxiv*. 2023.

Malik-Moraleda S, Jouravlev O, Taliaferro M, Mineroff Z, Cucu T, Mahowald K, Blank IA, Fedorenko E. Functional characterization of the language network of polyglots and hyperpolyglots with precision fMRI. *bioRxiv*. 2024: 2023 January 19. 524657.

Malik-Moraleda S, Ayyash D, Gallée J, Affourtit J, Hoffmann M, Mineroff Z, Jouravlev O, Fedorenko E. An investigation across 45 languages and 12 language families reveals a universal language network. *Nat Neurosci*. 2022:258(25):1014–1019.

Marslen-Wilson WD. Functional parallelism in spoken word-recognition. *Cognition*. 1987:25(1–2):71–102.

Mesgarani N, Cheung C, Johnson K, Chang EF. Phonetic feature encoding in human superior temporal gyrus. *Science*. 2014: 1006(6174):1006–1010.

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

Mollica F, Siegelman M, Diachek E, Piantadosi ST, Mineroff Z, Futrell R, Kean H, Qian P, Fedorenko E. Composition is the Core driver of the language-selective network. *Neurobiol Lang*. 2020:1(1):104–134.

Monaghan P, Shillcock RC, Christiansen MH, Kirby S. How arbitrary is language? *Philos Trans R Soc B Biol Sci*. 2014:369(1651):20130299.

Monti MM, Parsons LM, Osherson DN. Thought beyond language: neural dissociation of algebra and natural language. *Psychol Sci*. 2012:23(8):914–922.

Myers EB, Blumstein SE, Walsh E, Eliassen J. Inferior frontal regions underlie the perception of phonetic category invariance. *Psychol Sci*. 2009:20(7):895–903.

Norman-Haignere S, Kanwisher NG, McDermott JH. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *neuron*. 2015;88(6):1281–1296.

Nieto-Castañón A. *Handbook of functional connectivity magnetic resonance imaging methods in CONN*; Hilbert Press; 2020.

Nieto-Castañón A, Fedorenko E. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage*. 2012:63(3):1646–1669.

Norman-Haignere SV, Long LK, Devinsky O, Doyle W, Irobunda I, Merricks EM, Feldstein NA, McKhann GM, Schevon CA, Flinker A, et al. Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nat Hum Behav*. 2022:2022(3): 1–15.

O'Donnell TJ. *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press; 2015.

Okada K, Hickok G. Identification of lexical-phonological networks in the superior temporal sulcus using functional magnetic resonance imaging. *Neuroreport*. 2006:17(12):1293–1296.

Okada K, Matchin W, Hickok G. Phonological feature repetition suppression in the left inferior frontal gyrus. *J Cogn Neurosci*. 2017:30(10):1549–1557.

Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*. 1971;9(1):97–113.

Overath T, McDermott JH, Zarate JM, Poeppel D. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat Neurosci*. 2015:18(6):903–911.

Paulesu E, Frith CD, Frackowiak RSJ. The neural correlates of the verbal component of working memory. *Nature*. 1993:362(6418): 342–345.

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, Botvinick M. Fedorenko E (2018) toward a universal decoder of linguistic meaning from brain activation. *Nat Commun*. 2018:91(9): 1–13.

Perry LK, Perlman M, Winter B, Massaro DW, Lupyan G. Iconicity in the speech of children and adults. *Dev Sci*. 2018:21(3):12572.

Petersen SE, Fox PT, Posner MI, Mintun M, Raichle ME. Positron emission tomographic studies of the processing of single words. *J Cogn Neurosci*. 1989:1(2):153–170.

Pillay SB, Stengel BC, Humphries C, Book DS, Binder JR. Cerebral localization of impaired phonological retrieval during rhyme judgment. *Ann Neurol*. 2014:76(5):738–746.

Pimentel T, McCarthy AD, Blasi DE, Roark B, Cotterell R (2019) Meaning to form: measuring Systematicity as information. *ACL 2019 - 57th Annu Meet Assoc Comput Linguist Proc Conf*:1751–1764.

Pimentel T, Roark B, Cotterell R. Phonotactic complexity and its trade-offs. *Trans Assoc Comput Linguist*. 2020:8:1–18.

Pinker S. Rules of language. *Science (80-)*. 1991:253(5019):530–535.

Poldrack RA. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci*. 2006:10(2):59–63.

Poldrack RA, Wagner AD, Prull MW, Desmond JE, Glover GH, Gabrieli JDE. Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *NeuroImage*. 1999:10(1):15–35.

Price CJ. The anatomy of language: a review of 100 fMRI studies published in 2009. *Ann N Y Acad Sci*. 2010:1191(1):62–88.

Price CJ. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*. 2012:62(2):816–847.

Price CJ, Moore CJ, Humphreys GW, Wise RJS. Segregating semantic from phonological processes during reading. *J Cogn Neurosci*. 1997:9(6):727–733.

Pylkkänen L. The neural basis of combinatory syntax and semantics. *Science*. 2019:80-) 366(6461):62–66.

Quillen IA, Yen M, Wilson SM. Distinct neural correlates of linguistic and non-linguistic demand. *Neurobiology of Language* 2021;2(2):202–25.

Regev M, Honey CJ, Simony E, Hasson U. Selective and invariant neural responses to spoken and written narratives. *J Neurosci*. 2013:33(40):15978–15988.

Regev TI, Casto C, Hosseini EA, Adamek M, Ritaccio AL, Brunner P, Fedorenko E. Neural populations in the language network differ in the size of their temporal receptive windows. *bioRxiv*. 2023: 2022 December 30.522216.

Richardson H, Koster-Hale J, Caselli N, Magid R, Benedict R, Olson H, Pyers J, Saxe R. Reduced neural selectivity for mental states in deaf children with delayed exposure to sign language. *Nat Commun*. 2020, 2020:111(11):1–13.

Shain C, Paunov A, Chen X, Lipkin B, Fedorenko E. No evidence of theory of mind reasoning in the human language network. *Cerebral Cortex*. 2023;33(10):6299–6319.

Shain C, Kean H, Casto C, Lipkin B, Affourtit J, Siegelman M, Mollica F, Fedorenko E. Graded sensitivity to structure and meaning throughout the human language network. *bioRxiv*. 2023.

Scott TL, Perrachione TK. Common cortical architectures for phonological working memory identified in individual brains. *NeuroImage*. 2019:202:116096.

Scott TL, Gallée J, Fedorenko E. A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cogn Neurosci*. 2017:8(3):167–176.

Seghier ML, Lazeyras F, Pegna AJ, Annoni J-M, Zimine I, Ne Mayer E, Michel CM, Khateb A. Variability of fMRI activation during a phonological and semantic language task in healthy subjects. *Hum Brain Mapp*. 2004:23(3):140–155.

Shashidhara S, Mitchell DJ, Erez Y, Duncan J. Progressive recruitment of the Frontoparietal multiple-demand system with increased task complexity, time pressure, and reward. *J Cogn Neurosci*. 2019:31(11):1617–1630.

Sidhu DM, Pexman PM. Five mechanisms of sound symbolic association. *Psychon Bull Rev*. 2018:25(5):1619–1643.

Storkel HL. Learning new words: Phonotactic probability in language development. *J Speech, Lang Hear Res*. 2001:44(6):1321–1337.

Tyler LK, Marslen-Wilson WD, Randall B, Wright P, Devereux BJ, Zhuang J, Papoutsi M, Stamatakis EA. Left inferior frontal cortex and syntax: function, structure and behaviour in patients with left hemisphere damage. *Brain*. 2011:134(2):415–431.

Ullman MT. The declarative/procedural model: a neurobiological model of language learning, knowledge, and use. In: *Neurobiology of language*. Elsevier Inc.; 2015. pp. 953–968. https://www.sciencedirect.com/science/article/abs/pii/B9780124077942000766.

Vaden KI, Piquado T, Hickok G. Sublexical properties of spoken words modulate activity in Broca's area but not superior temporal cortex: implications for models of speech recognition. *J Cogn Neurosci*. 2011:23(10):2665–2674.

Vagharchakian L, Dehaene-Lambertz G, Pallier C, Dehaene S. A temporal bottleneck in the language comprehension network. *J Neurosci*. 2012:32(26):9089–9102.

Vigneau M, Beaucousin V, Hervé PY, Duffau H, Crivello F, Houdé O, Mazoyer B, Tzourio-Mazoyer N. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *NeuroImage*. 2006:30(4):1414–1432.

Vinckier F, Dehaene S, Jobert A, Dubus JP, Sigman M, Cohen L. Hierarchical coding of letter strings in the ventral stream: dissecting the inner Organization of the Visual Word-Form System. *Neuron*. 2007:55(1):143–156.

Vinson D, Jones M, Sidhu DM, Lau-Zhu A, Santiago J, Vigliocco G. Iconicity emerges and is maintained in spoken language. *J Exp Psychol Gen*. 2021:150(11):2293–2308.

Vitevitch MS, Luce PA. Probabilistic Phonotactics and Neighborhood activation in spoken word recognition. *J Mem Lang*. 1999:40(3):374–408.

Vitevitch MS, Luce PA. Increases in phonotactic probability facilitate spoken nonword repetition. *J Mem Lang*. 2005:52(2):193–204.

Vitevitch MS, Luce PA. Phonological Neighborhood effects in spoken word perception and production. *Annu Rev Linguist*. 2016:2(1):75–94.

Vitevitch MS, Luce PA, Pisoni DB, Auer ET. Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain Lang*. 1999:68(1–2):306–311.

Wedel A, Ussishkin A, King A. Crosslinguistic evidence for a strong statistical universal: phonological neutralization targets word-ends over beginnings. *Language (Baltim)*. 2019:95:e428–e446.

Weiss Y, Cweigenberg HG, Booth JR. Neural specialization of phonological and semantic processing in young children. *Hum Brain Mapp*. 2018:39(11):4334–4348.

Willems RM, Der Haegen L, Fisher SE, Francks C. On the other hand: including left-handers in cognitive neuroscience and neurogenetics. *Nat Rev Neurosci*. 2014, 2014:153(15):193–201.

Wilson SM, Saygin AP. Grammaticality judgment in aphasia: deficits are not specific to syntactic structures, aphasic syndromes, or lesion sites. *J Cogn Neurosci*. 2004:16(2):238–252.

Winter B, Perlman M, Perry LK, Lupyan G. Which words are most iconic? Interact stud Soc Behav Commun biol Artif Syst stud. *Soc Behav Commun Biol Artif Syst Stud*. 2017:18(3):443–464.

Woolnough O, Donos C, Rollo PS, Forseth KJ, Lakretz Y, Crone NE, Fischer-Baum S, Dehaene S. Tandon N (2020) spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nat Hum Behav*. 2020:53(5):389–398.

Xie X, Myers E. Left inferior frontal gyrus sensitivity to phonetic competition in receptive language processing: a comparison of clear and conversational speech. *J Cogn Neurosci*. 2018:30(3):267–280.

Yen M, DeMarco AT, Wilson SM. Adaptive paradigms for mapping phonological regions in individual participants. *NeuroImage*. 2019:189:368–379.

Zipf GK. *The psycho-biology of language*. George Routledge & Sons, Ltd; 1936.