The language network supports both lexical access and sentence generation during language production

Jennifer Hu*[1], Hannah Small*[1,2], Hope Kean[1,2], Atsushi Takahashi[2], Leo Zekelman[3], Daniel Kleinman[4], Elizabeth Ryan[5], Victor Ferreira[6], and Evelina Fedorenko[1,2,3]

[1]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA
[2]McGovern Institute for Brain Research, MIT, Cambridge, MA 02139, USA
[3]Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA 02138, USA
[4]Haskins Laboratories, 300 George Street, New Haven, CT 06511, USA
[5]St. George's Medical School, St. George's University, Grenada, West Indies
[6]Department of Psychology, UCSD, La Jolla, CA 92093, USA

* Equal contributors

Contributions:

Role/Author	JH*	HS*	HK	AT	LZ	DK	ER	VF	EF
Conceptualization	☑				☑	☑		☑	☑
Methodology	☑	☑			☑	☑	☑	☑	☑
Investigation (data collection)	☑	☑	☑	☑			☑		
Data curation	☑	☑							
Formal analysis	☑	☑							
Validation	☑	☑							
Visualization	☑	☑							
Software	☑	☑							
Resources				☑					
Writing – original draft	☑	☑							☑
Writing – review and editing			☑		☑	☑		☑	
Project administration									☑
Supervision									☑

Corresponding Authors
Jennifer Hu or Ev Fedorenko
jennhu@mit.edu / evelina9@mit.edu; 43 Vassar Street, Room 46-3037, Cambridge, MA, 02139

Acknowledgements

We would like to acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, and its support team. We thank former and current EvLab members (especially Josef Affourtit and Alvincé Pongos) for their help with data collection, the audience at the 2020 Neurobiology of Language conference, and members of the Fedorenko and Gibson labs for helpful comments and discussions. JH was supported by NIH award T32NS105587 and an NSF Graduate Research Fellowship. EF was supported by NIH R01 awards R01-DC016607 and R01-DC016950, and research funds from the Department of Brain and Cognitive Sciences and the McGovern Institute for Brain Research.

Abstract

A network of left frontal and temporal brain regions has long been implicated in language comprehension and production. However, because of relatively fewer investigations of language production, the precise role of this 'language network' in production-related cognitive processes remains debated. Across four fMRI experiments that use picture naming/description to mimic the translation of conceptual representations into words and sentences, we characterize the response of the language regions to production demands. In line with prior studies, sentence production elicited strong responses throughout the language network. Further, we report three novel results. First, we demonstrate that production-related responses in the language network are robust to output modality (speaking vs. typing). Second, the language regions respond to both lexical access and sentence-generation demands. This pattern implies strong integration between lexico-semantic and combinatorial processes, mirroring the picture that has emerged in language comprehension. Finally, some have previously hypothesized the existence of production-selective mechanisms given that syntactic encoding is a critical part of sentence production, whereas comprehension is possible even when syntactic cues are degraded or absent. Contrary to this hypothesis, we find no evidence of brain regions that selectively support sentence generation. Instead, language regions respond overall more strongly during production than during comprehension, which suggests that production incurs a greater cost for the language network. Together, these results align with the idea that language comprehension and production draw on the same knowledge representations, which are stored in the language-selective network and are used both to interpret linguistic input and generate linguistic output.

Introduction

Although the scientific enterprise of language neuroscience research began with a report about an aspect of language *production* (articulatory abilities; Broca, 1861), the field has been largely dominated by investigations of linguistic *comprehension*, plausibly because complex self-generated behaviors are notoriously challenging to study (e.g., Bock, 1996). As a result, many questions remain about the cognitive and neural mechanisms of language production.

At a broad level, a dissociation has been consistently observed between lower-level articulatory abilities and higher-level production abilities (lexical access and sentence generation). Our ability to produce sequences of speech sounds draws on a network of superior temporal and frontal areas, including an area in posterior left inferior frontal gyrus, in line with Broca's original report (e.g., Bohland & Guenther, 2006; Bouchard et al., 2013; Flinker et al., 2015; Fridriksson et al., 2016; Long et al., 2016; Basilakos et al., 2018; Guenther, 2016). This network does not appear to be sensitive to the meaningfulness of the productions, evidenced by its strong engagement in generation of even meaningless syllable sequences. In contrast, higher-level aspects of production—accessing words and putting them together into phrases/sentences—appear to draw on a different network of (more inferior) temporal and (more anterior) frontal areas (see, e.g., Indefrey & Levelt, 2004; Indefrey, 2011 for reviews). These areas appear to correspond to the language-selective network (e.g., Fedorenko et al., 2011) that also supports language comprehension (e.g., Fedorenko et al., 2010; Bautista & Wilson, 2016). Indeed, studies that have directly compared comprehension and production have observed overlap within these areas (e.g., Menenti et al., 2011; Silbert et al., 2014).

However, beyond this broad distinction between articulatory and higher-level aspects of production, the precise contributions of different language areas to lexical access and sentence generation remain unclear. Indefrey (2011) assigns distinct roles to different parts of the language network, but his proposal focuses on single-word retrieval, making it unclear how cognitive processes that are required to assemble words into phrases would fit into the picture: are they supported by the same areas, or do different/additional brain areas come into play? Some intracranial studies have reported relatively focal areas within the language network—in posterior temporal and inferior frontal cortex—wherein stimulation leads to impairments in phrase and sentence production but not in production of single words (e.g., Chang et al., 2018; Lee, Fedorenko et al., 2018; see Ding et al., 2020 for evidence from acute stroke patients). But in other studies, these same general areas have been linked to single-word production (posterior temporal cortex: e.g., Borovsky et al., 2007; Halai et al., 2017; inferior frontal cortex: e.g., Schnur et al., 2009; Corina et al., 2010; Kojima et al., 2013; Python et al., 2018).

To illuminate the contribution of the language-selective network to language production, we examined responses of the language areas—defined functionally by an extensively validated language 'localizer' (Fedorenko et al., 2010)—to word and sentence production. Adapting a paradigm that has proven fruitful in probing comprehension (e.g., Friederici et al., 2000; Humphries et al., 2006; Fedorenko et al., 2010, 2012a), we examine responses during the production of sentences, lists of words, and nonword sequences. Brain areas that support lexical retrieval should work harder (exhibit stronger responses) when words have to be accessed (word-list production) compared to a simple articulatory task (nonword production), given the additional computations required for word retrieval. Similarly, brain areas that support sentence generation should work harder when words have to be combined (sentence production) compared to retrieval of unrelated words, given that sentence production requires extra operations related to phrase-structure building. To ensure that the neural responses have to do with the hypothesized computations related to lexical access and sentence generation, we further assess their output-modality independence (speaking out loud vs. typing). Finally, we search across the brain for areas that may be selective for sentence production, given the different computational demands associated with sentence production vs. comprehension (e.g., Bock, 1995).

Materials and Methods

Participants

Forty-one individuals (age 18-31, mean 23.3 years; 28 (68.3%) females) from the Cambridge/Boston, MA community participated for payment across four fMRI experiments (n=15 in Experiment 1; n=14 in Experiments 2a and 2b; and n=12 in Experiment 3). All were native speakers of English. Of the thirty-two participants for whom handedness data were available, twenty-eight participants (87.5%) were right-handed, as determined by the Edinburgh Handedness Inventory (Oldfield, 1971) or self-report, two (6.25%) were left-handed, and two (6.25%) were ambidextrous (see Willems et al., 2014, for arguments for including non-right-handers in cognitive neuroscience experiments). Handedness data were not recorded for the remaining 9 participants. All but one participant showed typical left-lateralized language activations in the language localizer task; one (right-handed) participant in Experiment 3 showed right-lateralized language activations; we chose to include this participant to err on the conservative side. For Experiment 2, we recruited participants who could type without seeing the written output or the keyboard itself. All participants gave informed consent in accordance with the requirements of MIT’s Committee on the Use of Humans as Experimental Subjects (COUHES).

Design, materials, and procedure

Each participant completed a comprehension-based localizer task for the language network (Fedorenko et al., 2010) and a critical language production experiment. All but one participant (in Experiment 1) additionally completed a localizer task for the domain-general Multiple Demand (MD) network (Duncan, 2010, 2013). Because the MD network has been shown to be generally sensitive to task difficulty across domains (e.g., Duncan & Owen, 2000; Fedorenko et al., 2013; Hugdahl et al., 2015; Shashidhara et al., 2019; Assem et al., 2020), activity levels therein can be used to determine the relative difficulty levels of the different production conditions, to aid interpretation. Some participants also completed one or more tasks for unrelated studies. The scanning sessions lasted approximately two hours.

Language network localizer

The regions of the language network were localized using a task 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/). Briefly, participants read sentences and lists of unconnected, pronounceable nonwords in a blocked design. The sentences > nonwords contrast targets brain regions that that support high-level language comprehension. This contrast generalizes across tasks (e.g., Fedorenko et al., 2010; Scott et al., 2017; Ivanova et al., 2020) and presentation modalities (reading vs. listening; e.g., Fedorenko et al., 2010; Scott et al., 2017; Chen, Affourtit et al., 2021). All the regions identified by this contrast show sensitivity to lexico-semantic processing (e.g., stronger responses to real words than nonwords) and combinatorial syntactic and semantic processing (e.g., stronger responses to sentences and Jabberwocky sentences than to unstructured word and nonword sequences) (e.g., Fedorenko et al., 2010, 2012a, 2016, 2020;

Blank et al., 2016). More recent work further shows that these regions are sensitive to sub-lexical regularities (Regev et al., 2021), in line with the idea that this system stores our linguistic knowledge, which encompasses regularities across representational grains, from phonological schemas, to words, to constructions. Further, a network that closely corresponds to the one activated by the language localizer emerges from task-free (resting state) data (e.g., Braga et al., 2020).

Stimuli were presented one word/nonword at a time at the rate of 350-450ms (differing slightly between variants of the localizer; **Table SI-1)** per word/nonword. Participants read the materials passively and performed either a simple button-press or memory probe task at the end of each trial (included in order to help participants remain alert). The memory probe required the participant to indicate whether a given word was from the list of words/nonwords they had just read. Each participant completed two ~6 minute runs. In Experiments 1 and 2a/b, all participants completed the language localizer in the same session as the production experiment. In Experiment 3, 8 participants completed the language localizer in the same session as the production experiment and the remaining 4 participants completed the language localizer in an earlier scanning session.

MD network localizer

The regions of the MD network were localized using a spatial working memory task contrasting a harder condition with an easier condition (e.g., Fedorenko et al., 2011, 2013; Blank et al., 2014). The hard > easy contrast targets brain regions engaged in cognitively demanding tasks.

Fedorenko et al. (2013) have established that the regions activated by this task are also activated by a wide range of other demanding tasks (see also Duncan and Owen, 2000; Fedorenko et al., 2013; Hugdahl et al., 2015; Shashidhara et al., 2019; Assem et al., 2020).

On each trial (8 s), participants saw a fixation cross for 500 ms, followed by a 3 x 4 grid within which randomly generated locations were sequentially flashed (1 s per flash) two at a time for a total of eight locations (hard condition) or one at a time for a total of four locations (easy condition). Then, participants indicated their memory for these locations in a two-alternative, forced-choice paradigm via a button press (the choices were presented for 1,000 ms, and participants had up to 3 s to respond). Feedback, in the form of a green checkmark (correct responses) or a red cross (incorrect responses), was provided for 250 ms, with fixation presented for the remainder of the trial. Hard and easy conditions were presented in a standard blocked design (4 trials in a 32 s block, 6 blocks per condition per run) with a counterbalanced order across runs. Each run included four blocks of fixation (16 s each) and lasted a total of 448 s. Each participant completed two runs, except for one subject in Experiment 2a/b, who completed one run. In Experiment 1, all participants who completed the MD localizer (n=14/15) did so in the same session as the production experiment. In Experiment 2a/b, 11 participants completed the MD localizer in the same session as the production experiment and the remaining 3 participants completed the MD localizer in an earlier scanning session. In Experiment 3, 9 participants completed the MD localizer in the same session as the production experiment and the remaining 3 participants completed the MD localizer in an earlier scanning session.

Like the language localizer, the MD localizer has been extensively validated, and a network that closely corresponds to the one activated by the MD localizer emerges from task-free (resting state) data (e.g., Braga et al., 2020; Assem et al., 2020).

General approach for language production tasks

Tapping mental computations related to high-level language production—including both lexical access and combining words into phrases and sentences—is notoriously challenging because linguistic productions originate from internal conceptual representations (e.g., Levelt, 1989; Bock, 1996). These representations are difficult to probe and manipulate without sacrificing ecological validity. Given the many open questions that remain about how language production is implemented in the mind and brain, and the need for careful comparisons (critical for interpretability), we opted for a controlled experimental approach. In particular, building on a strong foundation of behavioral work on language production, we used pictorial stimuli to elicit object labels and sentence-level linguistic descriptions.

Experiment 1 (Speaking)

Design. Participants were presented with a variety of visual stimuli across six conditions. In the two critical language production conditions—sentence production and word-list production—participants were instructed to speak out loud, but to move their heads as little as possible. In the sentence production (SProd) condition, which is the closest to reflecting the language production demands of everyday life, where we often communicate event-level descriptions using phrases and sentences, participants viewed photographs of common events involving humans, animals, and inanimate objects (**Figure 1a-i, 1b-i**) and were asked to produce a description of the event (e.g., “The girl is smelling a flower”). This condition targets sentence-level production planning and execution, which includes a) retrieving the words for the entities/objects and actions, and b) combining them into an utterance, including ordering the words and implementing the relevant syntactic agreement processes. In the word-list production (WProd) condition, participants viewed groups of 2, 3, or 4 photographs of inanimate objects (**Figure 1a-ii, 1b-ii**) and were asked to name each object in the set (e.g., “accordion, ladder, apple”). The number of objects in each group (2-4) matched the number of content words in the target productions in the SProd condition. This condition targets word-level production planning and execution. To isolate the mental processes related to single-word production, photographs were manually grouped in a way that minimized semantic associations between the objects, to prevent participants from unintentionally forming phrases/clauses with the retrieved words.

The experiment also included two control conditions: low-level (nonword-list) production and semantic judgments about visual events. In the nonword-list production (NProd) condition, participants viewed lists of 4 monosyllabic nonwords (**Figure 1a-iii**) and were asked to say them out loud (e.g., “blolt, sloal, sneaf, tworce”). The nonwords obeyed the phonotactic constraints of English and were selected to be sufficiently distant from phonologically neighboring words. This condition targets low-level articulatory planning and execution and was included in order to isolate aspects of language production related to lexical access and sentence generation. In the visual event semantics (VisEvSem) condition, participants viewed photographs of events (as in SProd) and were asked to indicate whether the depicted event takes place indoors or outdoors (a relatively high-level judgment that requires visual event perception and also draws on world knowledge) via a two-choice button box (**Figure 1a-iv**). This condition targets visual and conceptual processing of events and was included to ensure that responses to the SProd condition, which uses these pictorial stimuli, were not due to these cognitive processes.

Finally, the experiment included two reading comprehension conditions: sentence comprehension and word-list comprehension. In both conditions, participants were instructed to read the stimuli silently (as in the language localizer). In the sentence comprehension (SComp) condition, participants viewed short sentences describing common events, similar to the events depicted in the photographs used in the SProd condition (**Figure 1a-v**) and were asked to read them (e.g., “A woman is playing the harp”). This condition targets sentence-level comprehension processes, including lexico-semantic and combinatorial (syntactic and semantic) processes. In the word-list comprehension (WComp) condition, participants viewed lists of 2, 3, or 4 object names (**Figure 1a-vi**) and were asked to read them (e.g., “arm, deer, bubble”). This condition targets word-level comprehension. As in the WProd condition, object names were grouped in a way that minimized semantic associations. These conditions were included to directly compare the responses to content-matched sentences and word lists across production and comprehension as relevant to the question of production-selective mechanisms.

Materials. To obtain the event photographs for the SProd and VisEvSem conditions, we first manually selected 400 images clearly depicting everyday events from the Flickr30k dataset (Young et al., 2014). We then ran a norming study on Amazon.com’s Mechanical Turk to identify the stimuli that would elicit the most consistent linguistic descriptions across participants. On each trial, participants viewed a single photograph and were given the instructions "Please provide a one-sentence description of what is happening in the photo." They were able to type freely in a textbox below the image and could only proceed to the next trial after submitting a non-empty response. We recruited n=30 participants for each of the 400 images, and each participant produced descriptions for 100 images.

To analyze the resulting 12,000 responses, we used the Python spaCy natural language processing library (Honnibal et al., 2020) to parse each production into the subject noun phrase (NP), verb phrase (VP), subject NP head, and VP head. After manually cleaning the parses for consistency, we computed three metrics for each photograph: (1) the number of unique responses in each of the parsed categories, (2) the number of unique lemmas for the single-word parsed categories (subject NP head and VP head), and (3) the standard deviation of the number of tokens per production. We then obtained a 'linguistic variability' score by summing these three values for each image and chose the 200 photographs with the lowest scores. Finally, we hand-selected 128 from these 200 to maximally cover a range of objects and actions. These photographs were used in the SProd and VisEvSem conditions, and the associated sentence descriptions (the most frequently used description for each photograph) were used in the SComp condition.

For the WProd and WComp conditions, we wanted to use materials that would be semantically (and lexically) similar to the ones used in the SProd and SComp conditions. As a result, to obtain the object photographs for the WProd condition, we first identified between 2 and 4 words in each of the 128 sentence descriptions that referred to inanimate objects (we avoided animate entities like ‘a man’ or ‘a woman’ because in the setup that we used, with multiple objects presented at once, we wanted to avoid the possibility of participants constructing event-level representations). For example, from the description “A man is playing saxophone in a park” we selected ‘saxophone’, and from the description “A man is sitting on a bench reading the newspaper” we selected ‘bench’ and ‘newspaper’. This resulted in a total of 120 words. Next, we selected images of each object from the THINGS database (Hebart et al., 2019) as well as a repository of license-free stock photographs. In those images, each object is presented on a neutral but naturalistic background, which isolates the object from possibly associated events or concepts. We generated all possible groups of 2-, 3-, and 4-object images, and then took a random sample of 40 2-object, 80 3-object, and 40 4-object groups, as there was an average of 3 content words in our target sentence productions. We then manually selected the final 128 object groups by discarding groups with semantically related objects and ensuring that each object appeared 1-3 times. The associated words (grouped in the same way) were used in the WComp condition. The order of objects and words was randomized within each group during presentation.

Finally, the nonwords for the NProd condition were selected from the ARC Nonword Database (Rastle et al., 2002). We began by selecting all the monomorphemic syllables involving orthographically existing onsets, bodies, and legal bigrams. We then obtained the final set of 256 nonwords by filtering for low numbers of onset and phonological neighbors in order to minimize the likelihood of these nonwords priming real words. These 256 nonwords were then randomly distributed into 64 groups of 4.

Procedure. To ensure that the same event or object group would not appear in both a production condition and its corresponding comprehension condition for any given participant, we distributed the materials in the SProd, WProd, SComp, and WComp conditions (i.e., 128 event images, 128 corresponding target sentences, 128 object group images, and 128 corresponding target word lists) into two experimental lists. To do so, we assigned a unique number 1-128 to each event and object group, such that event image *x* corresponds to sentence *x*, and object group *x* corresponds to word list *x*. These numbers were assigned such that sets 1-64 and 65-128 were each semantically diverse (e.g., two images of a person playing a musical instrument were assigned to different sets). Furthermore, the 2-, 3-, and 4-object groups were evenly distributed across the two sets (1-64 and 65-128). Finally, this numbering was used to create two lists. In List 1, event images 1-64 in SProd appeared with sentences 65-128 in SComp, and similarly object group images 1-64 in WProd appeared with word lists 65-128 in WComp. And in List 2, event/object group images 65-128 in SProd/WProd appeared with sentences/word lists 1-64 in SComp/WComp. The materials for the NProd condition were identical across lists, and the materials for the VisEvSem condition were the same as the SProd materials in that list.

The materials in each condition (and each list, where relevant) were grouped into 16 blocks of 4 trials each; this was done separately for each participant. (Note that although we had enough materials to have 16 blocks per condition, we ended up presenting 12 blocks per condition for any given participant because—based on pilot participants—this number of blocks per condition gave us sufficient power to elicit clear between-condition differences.) Each block was preceded by instructions, which told the participants what they would be doing in the trials to come: “Describe the event out loud” for SProd, “Name the objects out loud” for WProd, “Say the nonwords out loud” for NProd, “Inside (=1) or outside (=2)?” for VisEvSem, “Read the sentence silently" for SComp, and "Read the words silently" for WComp. The instructions remained on the screen (in small font in the bottom left corner of the screen) throughout the block to minimize the demands associated with holding onto the instructions and to help participants in case they missed the block-initial instructions screen. Each trial lasted 3 sec and consisted of an initial fixation cross (0.2 sec) and stimulus presentation (2.8 sec). In the SProd and VisEvSem conditions, the stimulus was a single event picture; in the WProd condition, the stimulus was a set of 2-4 object pictures (presented all at once); in the NProd condition, the stimulus was a set of 4 nonwords (presented all at once); in the SComp condition, the stimulus was a sentence (presented all at once); and in the WComp, the stimulus was a set of 2-4 words (presented all at once) (see **Figure 1a**). The block-initial instructions were presented for 2 sec. Thus, each block lasted 14 sec (2 sec instructions and 4 trials 3 sec each).

The total of 72 experimental blocks (12 blocks * 6 conditions) were distributed into 6 sets, corresponding to runs, of 12 blocks each (2 blocks per condition). Each run additionally included 3 fixation blocks of 12 sec each: one at the beginning of the run, one after the first six experimental blocks, and one at the end. Thus, each run consisted of 12 experimental blocks of 14 sec each and 3 fixation blocks of 12 sec each, lasting a total of 204 sec (3 min 24 sec). Each participant completed 6 runs. The order of conditions was palindromic within each run and varied across runs and participants.

Prior to entering the scanner, participants were provided with printed instructions and were guided through sample items that mimicked the experimental stimuli. The experimental script with all the materials is available at GitHub: https://github.com/jennhu/LanguageProduction.

Experiments 2a (Speaking) and 2b (Typing).

Design and materials. For Experiment 2, each participant performed two experiments: Experiment 2a served to replicate Experiment 1, and Experiment 2b served to generalize the results from Experiment 1 to another output modality. The design of Experiment 2b was identical to that of Experiment 1, except that in the production conditions (the two critical conditions—SProd and WProd— and the NProd control condition), participants were asked to type their responses on a scanner-safe keyboard (described below) instead of speaking them out loud. For the control VisEvSem condition, participants were asked to type their answers (1 or 2) on the keyboard instead of the button box. The two critical production conditions target the same mental processes as in Experiment 1, and the control NProd condition targets low-level hand motor planning and execution, thus helping isolate higher-level aspects of language production. For any given participant, different experimental lists (see Experiment 1 for details) were used for Experiments 2a and 2b.

Procedure. The procedure for Experiment 2a was identical to that of Experiment 1, and the procedure for Experiment 2b only differed in the trial timing for the three production conditions (SProd, WProd, and NProd). In particular, for these conditions, trial duration was increased from 3 to 7 sec (0.2 sec fixation and 6.8 stimulus presentation) given that typing takes longer than speaking (especially when typing in an unusual position, as described below). Each run therefore consisted of 12 experimental blocks (6 were 14 sec each, as in Experiments 1 and 2a, and 6 were 23 sec each) and 3 fixation blocks of 12 sec each, lasting a total of 300 seconds (5 min). The on-screen instructions for the production conditions were also adjusted to reflect the difference in output modality: “Type a description of the event” for SProd, “Type the names of the objects” for WProd, and “Copy the nonwords (typing)” for NProd. Each participant completed 4-6 runs (for a total of 8-12 blocks per condition) of each of Experiments 2a and 2b. The order of experiments was counterbalanced across participants.

To collect the typed responses, we built a custom MR-safe wireless keyboard. We purchased an off-the shelf wireless keyboard (Inland model ic210) and removed all the ferrous mechanical parts, such as the case screws and the steel wires used to stabilize the wide keys (Shift, Return, and space keys). We then replaced the highly ferrous alkaline AA battery and pulse width modulated step-up voltage regulator with a lithium ion polymer (LiPo) battery and a linear low-drop out voltage regulator. The keyboard uses silicon dome switches and flexible conductive traces that were not found to be ferrous. The wireless USB receiver was plugged into the MRI suite’s penetration panel through a USB to DB9 filter to prevent the introduction of RF interference into the MR images. The absence of RF interference introduced by the keyboard was confirmed by collecting time series of BOLD scans with and without the presence of the keyboard and keys being pressed during these scans and calculating the pixel-by pixel temporal SNR (tSNR) on a static quality assurance phantom.

During the experiment, the keyboard was placed directly on the participant’s abdomen or on a small non-ferrous platform placed on their abdomen, so they could quite comfortably type while lying in the scanner (akin to working on one’s laptop in bed); however, they were unable to see the output of their typing or their own keystrokes. Participants were given a chance to practice typing prior to the experiment to get accustomed to the setup and the keyboard layout. We collected and monitored the participants' productions on a computer outside the scanning room. The experimental script with all the materials is available at GitHub: https://github.com/jennhu/LanguageProduction.

Experiment 3 (Speaking).

Design. Experiment 3 served to conceptually replicate the findings from the critical production conditions (SProd and WProd) in Experiments 1 and 2a while generalizing the results to a new set of materials. The design was identical except that in the word-list production (WProd) condition, participants always viewed groups of 3 object photographs (cf. 2, 3, or 4 object photographs in Experiments 1 and 2a), and they were asked to name each object in the set with an indefinite article (e.g., "a necklace, a pumpkin, a hammer"), which includes some basic phrase-level combinatorial processing in addition to lexical retrieval. The experiment included two other conditions that are not directly relevant to the current investigation and are therefore not discussed.

Materials. The materials were selected from the publicly available images in the Google Images database and consisted of 96 event images for the SProd condition (**Figure 1b-i**), and 288 object images for the WProd (**Figure 1b-ii**) condition. The event photographs were similar in style to those used in Experiments 1 and 2a, but were more semantically diverse, including not only humans interacting with inanimate objects (as most events in Experiments 1 and 2a), but also humans interacting with other humans, and humans interacting with animals. The object photographs were also similar in style to those used in Experiments 1 and 2a, but did not include any background, and were also more semantically diverse, including not only inanimate objects, but also humans (where the occupation of the person is clear: e.g., a chef, a juggler, a ballerina, etc.) and animals. As in Experiments 1 and 2a, the object photographs were grouped in a way that minimized semantic associations between the objects.

Procedure. The materials in each condition were grouped into 24 blocks of 4 trials each; this was done separately for each participant. (The materials were further divided into two experimental lists of 12 blocks per condition.) Each block was preceded by instructions, which told the participants what they would be doing in the trials to come: “Describe the events” for SProd, and “Name the objects” for WProd. Each trial lasted 4 sec and consisted of an initial fixation cross (0.25 sec) and stimulus presentation (3.75 sec). In the SProd condition, a trial consisted of a single event picture, and in the WProd condition, a trial consisted of three object pictures (presented all at once in a triangular configuration) (see **Figure 1b**). Participants were instructed to describe the events with complete sentences (e.g., “The woman is tossing a frisbee”) and to name the objects with indefinite determiners (e.g., “a necklace, a pumpkin, a hammer”). The block-initial instructions were presented for 2 sec. Thus, each block lasted 18 sec (2 sec instructions and 4 trials 4 sec each).

The total of 48 experimental blocks in each list (12 blocks * 4 conditions, 2 of which are of interest to the current study) were distributed into 4 sets, corresponding to runs, of 12 blocks each (3 blocks per condition). Each run additionally included 4 fixation blocks of 18 sec each: one at the beginning of the run, and one after each set of four experimental blocks. Thus, each run consisted of 12 experimental blocks of 18 sec each and 4 fixation blocks of 18 sec each, lasting a total of 288 sec (4 min 48 sec). Eleven participants completed 4 runs (for a total of 12 blocks per condition) and one participant completed 2 runs (for a total of 6 blocks per condition). The order of conditions was palindromic within each run and varied across runs and participants. Prior to entering the scanner, participants were provided with printed instructions and were guided through sample items that mimicked the experimental stimuli. The experimental script with all the materials is available at GitHub: https://github.com/jennhu/LanguageProduction.

fMRI data acquisition, preprocessing, and first-level modeling

Data acquisition

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: thirty-one 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 (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.

Preprocessing

Data preprocessing was carried out with SPM12 (using default parameters, unless specified otherwise) and supporting custom MATLAB scripts. Preprocessing of functional data included motion correction (realignment to the mean image using 2nd-degree b-spline interpolation), normalization into a common space (Montreal Neurological Institute (MNI) template) (estimated for the mean image using trilinear interpolation), resampling into 2 mm isotropic voxels, smoothing with a 4 mm FWHM Gaussian filter and high-pass filtering at 128 s.

First-level modeling

For both the language localizer task and the critical experiments, a standard mass univariate analysis was performed in SPM12 whereby a general linear model (GLM) estimated, for each voxel, the effect size of each condition in each experimental run. These effects were each modeled with a boxcar function (representing entire blocks/events) convolved with the canonical Hemodynamic Response Function (HRF). The model also included first-order temporal derivatives of these effects, as well as nuisance regressors representing entire experimental runs, offline-estimated motion parameters, and outlier time points (i.e., timepoints where the scan-to-scan differences in global BOLD signal were above 5 standard deviations, or where the scan-to-scan motion was above 0.9mm).

Definition of the language network functional regions of interest (fROIs)

For each participant, we defined a set of language fROIs using group-constrained, subject-specific localization (Fedorenko et al., 2010). In particular, each individual map for the *sentences* > *nonwords* contrast from the language localizer was intersected with a set of six binary masks. These masks (**Figure 2a;** available at https://evlab.mit.edu/funcloc/) 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. These masks covered the fronto-temporal language network in the left hemisphere. 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. (Note that we here included the language fROI within the angular gyrus (AngG) even though this fROI consistently patterns differently from the rest of the language fROIs in its functional response profile and patterns of functional correlations. In other recent papers, we have started excluding the AngG language fROI to focus on the coherent set of five language fROIs. However, we chose to include it here given the importance of the 'dorsal stream'—white matter tracts of the arcuate and/or superior longitudinal fasciculus that connect posterior-most temporal/parietal language areas and inferior frontal language areas—in language production (e.g., Hickok & Poeppel, 2004; Fridriksson et al., 2016).)

Definition of the MD network fROIs

For each participant, we defined a set of MD fROIs using group-constrained, subject-specific localization (Fedorenko et al., 2010). Each individual map for the *hard* > *easy spatial working* *memory* contrast from the MD localizer was intersected with a set of twenty binary masks (10 in each hemisphere). These masks (**Figure 3a;** available at https://evlab.mit.edu/funcloc/) were derived from a probabilistic activation overlap map for the same contrast in a large set of participants (n=197) using watershed parcellation. The masks covered the frontal and parietal components of the MD network (Duncan, 2010, 2013) bilaterally. Within each mask, a participant-specific MD fROI was defined as the top 10% of voxels with the highest t-values for the localizer contrast.

Analyses

All analyses were performed with linear mixed-effects models using the “lme4” package in R (version 1.1.26; Bates et al., 2015) with *p*-value approximation performed by the “lmerTest” package (version 3.1.3; Kuznetsova et al., 2017) and effect sizes (Cohen’s *d*) estimated by the “EMAtools” package (version 0.1.3; Kleiman, 2017).

Validation of the language and MD fROIs

To ensure that the language and MD fROIs behave as expected (i.e., show a reliable localizer contrast effect), we used an across-runs cross-validation procedure (e.g., Nieto-Castañón & Fedorenko, 2012). In these analyses, 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) 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. As expected, the language fROIs showed a robust *sentences* > *nonwords* effect (*p*s$<10^{-8}$, |*d*|s>2.05; *p*-values corrected for the number of fROIs using the False Discovery Rate (FDR) correction; Benjamini & Yekutieli, 2001), and the MD fROIs showed a robust *hard* > *easy spatial working memory* effect (*p*s$<10^{-4}$, |*d*|s>1.83). For subjects with only a single run of a given localizer, the activation maps were visually inspected to ensure they look as expected.

Critical analyses

To estimate the responses in the language fROIs (and MD fROIs for some of the analyses) to the conditions of the critical experiments, the data from all the runs of the language (or MD) localizer were used to define the fROIs, and the responses to each condition were then estimated in these regions. Statistical analyses were performed on these extracted PSC values.

To characterize the responses in the language network to language production, we asked three questions (Questions 1-3 below). First, we asked whether sentence production elicits responses in the language regions, as expected based on prior reports (e.g., Menenti et al., 2011; Silbert et al., 2014). Second, we asked whether responses to sentence production generalize across output modality (speaking vs. typing). And third, we asked whether language regions are sensitive to both lexical access and sentence generation—two core aspects of high-level language production—and whether different regions may be more sensitive to one or the other. Further, we asked (Question 4 below) whether *any brain regions*—within the boundaries of the language network or in the rest of the brain—support sentence generation but not sentence comprehension given the critical role of syntactic information in sentence production, in contrast to comprehension, which is possible even when syntactic cues are degraded or absent (e.g., Ferreira et al., 2002; Levy, 2008; Levy et al., 2009; Gibson et al., 2015).

For each relevant contrast (as described in detail below), we used two types of linear mixed-effect regression models: i) the language network model, which examined the language network as a whole and served as our primary analysis; and ii) the individual language fROI model, which examined each language fROI separately, to paint a more complete picture. 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 comprehension over non-linguistic processes (e.g., Fedorenko et al., 2011; Fedorenko & Blank, 2020) and with respect to their role in lexico-semantic and syntactic processing in comprehension (e.g., Fedorenko et al., 2010, 2012, 2016, 2020; Bautista & Wilson, 2016; Blank et al., 2016) (the AngG language fROI sometimes patterns a little differently from the rest of the language fROIs, as noted above); and b) exhibit strong inter-region correlations in both their activity during naturalistic cognition paradigms (e.g., Blank et al., 2014; Paunov et al., 2019 Braga et al., 2020) and in key functional markers, like the strength of response or the extent of activation in response to language stimuli (e.g., Mahowald & Fedorenko, 2016; Mineroff, Blank et al., 2018; Lipkin et al., in prep.). However, to allow for the possibility that language regions differ in their response to language production, and because for some questions we are explicitly interested in potential differences among the language regions, we supplement the network-wise analyses with the analyses of the six language fROIs separately.

For the network-wise analyses, we fit a linear mixed-effect regression model, predicting the level of BOLD response in the language fROIs in the contrasted conditions. The model included a fixed effect for condition and random intercepts for fROI and participant.

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

For each of the six language fROIs, we fit a linear mixed-effect regression model, predicting the level of BOLD response in the target language fROI in the contrasted conditions. The model included a fixed effect for condition and a random intercept of participant. The results were FDR-corrected for the six ROIs.

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

1. Does sentence production elicit a response in the language network?

To test whether language regions respond during sentence production, we used three contrasts. First, we compared the responses to the spoken sentence production condition (SProd in Experiments 1, 2a, and 3) against the fixation baseline. Second, we compared the responses to the sentence production condition against the response to the nonword strings condition from the language localizer—an unstructured and meaningless linguistic stimulus (we chose the nonwords condition from the language localizer rather than low-level production condition, NProd, because the latter was not included in Experiment 3, and we wanted to make the analyses parallel across experiments). And third, in Experiments 1 and 2a, we further compared the responses to the sentence production condition against the response to the visual event semantics condition (VisEvSem). A brain region that supports sentence production should exhibit a response during the SProd condition that falls above both the fixation baseline and the nonword strings condition. Further, if that brain region responds to production demands rather than the visual/conceptual processing associated with the event pictures, it should also respond more strongly during sentence production than during a semantic task on the same pictures.

2. Does the language network's response to sentence production generalize across output modality?

To test whether language regions respond to sentence production across modalities, we compared the responses to the typed sentence production condition (SProd in Experiment 2b) against the fixation baseline, against the response to the nonword strings condition from the language localizer, and against the response to the visual event semantics condition. We further directly compared the responses to the SProd spoken and SProd typed conditions in Experiment 2 to test whether different language regions, or the language network overall, show a preference for spoken vs. typed responses. (In a reality-check analysis, we also searched for regions that responded to lower-level production demands related to articulation and typing using a whole-brain group-constrained, subject-specific approach (Fedorenko et al., 2010) described in more detail in Section 4 below.)

3. Does the language network respond to both lexical access and sentence generation?

To characterize the responses of the language regions to lexical access, we compared the responses to the word-list production condition (WProd) in Experiments 1, 2a, and 2b against the low-level nonword production condition (NProd); and to characterize the responses to sentence generation, we compared the responses to the sentence production condition (SProd) in Experiments 1, 2a, 2b, and 3 against the word-list production condition. As discussed in the Introduction, a brain region that supports lexical access should respond more strongly during the

WProd condition than the NProd condition, matched for articulation demands. A brain region that supports sentence generation (including ordering the words and selecting the right morpho-syntactic word forms) should respond more strongly during the SProd condition (which requires both lexical access and phrase/sentence construction) than the WProd condition (which only requires lexical access).

Prior to the neural analyses above, we examined the sentence productions (typed responses) in Experiment 2b (spoken productions were not recorded). In particular, we wanted to know how often full noun phrases and tensed verb phrases were used in the productions, given that in past studies in our and other groups, event pictures have sometimes elicited 'newspaper headline'-style descriptions, in spite of the instructions to produce complete sentences (e.g., "girl smelling flower"/ "smelling a flower", instead of "a/the girl is smelling a flower"). Note that these versions are not ungrammatical; they just use an alternative, simplified syntactic frame. Generating such descriptions still requires operations beyond single-word retrieval (e.g., the words must be ordered appropriately and the right morpho-syntactic form of each word must be selected), but in at least some cases, agreement processes may not be required. Production strategies varied across participants, ranging from exclusive reliance on simplified syntax to producing complete sentences most of the time (see **Figure SI-1** for details). Importantly, the propensity to produce well-formed sentences was not predictive of the magnitude of the SProd>WProd effect across participants (Pearson r=-0.17, p=0.56), suggesting that this effect was more strongly driven by semantic composition / ordering the words.

In a control analysis, we asked whether the SProd condition might be eliciting stronger responses than the WProd and NProd conditions because it is more cognitively demanding. To test this, we examined the responses to these three conditions in a set of brain regions that have been previously established to be robustly sensitive to general cognitive effort across domains: the regions of the fronto-parietal Multiple Demand (MD) network (Duncan, 2010, 2013; Fedorenko et al., 2013; Hudgdahl et al., 2015; Shashidhara et al., 2019; Assem et al., 2020). This network is functionally distinct from the language network (see Fedorenko & Blank, 2020 for a review), and appears to respond during linguistic tasks only in the presence of external task demands, at least for language comprehension (Diachek, Blank, Siegelman et al., 2020; Fedorenko & Shain, submitted).

4. Do any brain regions selectively support sentence generation?

Finally, we asked whether any brain regions—including outside the language network—showed selective responses to computations related to sentence generation relative to i) sentence comprehension, and ii) lexical access in production. To test this, we used a whole-brain group-constrained, subject-specific approach (Fedorenko et al., 2010) to search for brain regions that respond more strongly during the SProd condition than each of the SComp and WProd conditions. This approach is akin to the traditional whole-brain random-effects analysis (Holmes & Friston, 1998), but is more statistically powerful and robust given that it a) takes into account inter-individual variability in the precise locations of functional areas, and b) has built into it an across-runs cross-validation procedure to ensure that the regions that emerge show replicable responses over time. Using the data from Experiments 1 and 2a, we created for each participant a whole-brain map that represented a conjunction of contrasts (SProd>SComp thresholded at p<0.001 uncorrected level, and SProd>WProd thresholded at p <0.05 uncorrected level; note that liberal thresholds are permissible in this approach given the across-runs cross-validation, as noted above). Each participant's map was binarized, with 1s corresponding to voxels that show reliable effects for both contrasts above, and 0s otherwise. These individual maps were then overlaid, and watershed parcellation was performed, as described in Fedorenko et al. (2010; see also Julian et al., 2012), to search for areas that would contain supra-threshold voxels in at least half of the participants. The resulting regions were then used as masks to define the individual fROIs using the same two contrasts, selecting the top 10% of voxels based on the *t*-values for each contrast and taking the intersection of those voxel sets (the n% approach allows for the definition of the fROIs in each individual). Finally, across-runs cross-validation was used to estimate the responses to the critical conditions (SProd, SComp, and WProd) in these individually defined fROIs and to test for the replicability of the SProd>SComp and SProd>WProd contrasts.

Results

1. In line with past studies (e.g., Menenti et al., 2011; Silbert et al., 2014), spoken sentence production elicited a robust response in the language network across the three experiments. This response was stronger than the fixation baseline (Experiment 1: d=2.241, p<0.001; Experiment 2a: d=1.527, p<0.001; Experiment 3: d=1.731, p<0.001), the nonword reading control condition from the language localizer (Experiment 1: d=1.832, p<0.001; Experiment 2a: d=1.054, p<0.001; Experiment 3: d=1.587, p<0.001), and visual event semantic processing (Experiment 1: d=2.081, p<0.001; Experiment 2a: d=1.037, p<0.001) (**Figure 2a**, **Table 1**). (These effects also held in all (Experiments 1 and 3) or most (Experiment 2a) individual language fROIs (**Figure 2b-g**, **Table 1**).)

2. Similarly, typed sentence production (SProd-typed, Experiment 2b) elicited a stronger response in the language network than fixation (d=1.027, p<0.001), nonword reading (d=0.498, p=0.003), and visual event semantic processing (d=0.576, p<0.001) (**Figure 2a**, **Table 2**). (The SProd-typed>Fixation and SProd-typed>VisEvSem effects also held for most of the individual language fROIs; the SProd-typed>Nonwords effect showed a positive numerical trend in five of the six fROIs (**Figure 2b-g**, **Table 2**).)

Our reality-check analysis successfully recovered sets of brain regions in the premotor and motor cortex that selectively respond to articulation (**Figure SI-2a**) and to typing (**Figure SI-2b**) relative to fixation. In contrast to the language fROIs, these lower-level brain areas did not strongly discriminate among the articulation / typing conditions based on their linguistic content, and the articulation-selective regions showed a pattern opposite to that observed in the language network, with strongest responses to nonword production and lowest responses to sentence production.

Finally, at the network level, the SProd-spoken condition (Experiment 2a) elicited a stronger response than the SProd-typed condition (Experiment 2b) (d=0.459, p=0.006) (**Figure 2a, Table 2**). (This effect also showed a positive numerical trend in all six fROIs (**Figure 2b-g**).)

3. The language network responded to both lexical access and sentence generation. At the network level, word production (WProd) elicited a stronger response than nonword production (NProd) in Experiments 1 (d=0.333, p=0.037) and 2a (d=0.376, p=0.023) (**Figure 2, Table 3**). (This effect was not observed in Experiment 2b, where typed production was used (d=-0.105, p=0.524; we speculate on possible reasons in the Discussion) (**Figure 2, Table 3**). Similarly, sentence production (SProd) elicited a stronger response than word production (WProd) in all experiments, across spoken and typed modalities (Experiment 1: d=0.668, p<0.001; Experiment 2a: d=0.356, p=0.032; Experiment 2b: d=0.830, p<0.001; Experiment 3: d=0.609, p<0.001) (**Figure 2, Table 3**). (Similar patterns held in the individual language fROIs, with the SProd>WProd effect being generally stronger than the WProd>NProd effect, and coming out as reliable in all or most individual fROIs in Experiments 1, 2b and 3 (**Figure 2b-g**, **Table 3**).)

To control for the possibility that the language network's strong response to sentence production (SProd) is simply due to sentence production being a generally more difficult task than word production (WProd) and nonword production (NProd), we performed a control analysis in the

MD network, which is robustly sensitive to task demands across domains (e.g., Duncan & Owen, 2000; Fedorenko et al., 2013; Hugdahl et al., 2015; Shashidhara et al., 2019; Assem et al., 2020). The MD network responded more strongly to the WProd condition than to the SProd condition across experiments (Experiment 1: d=0.692, p<0.001; Experiment 2a: d=0.612, p<0.001; Experiment 2b: d=0.225, p=0.01; Experiment 3: d=0.353, p<0.001), providing evidence that WProd is, in fact, more cognitively demanding than SProd (**Figure 3, Table 4**). The NProd condition elicited a reliably stronger response than the SProd condition in Experiments 2a (d=0.212, p=0.015) and 2b (d=0.228, p=0.009), and was similar in magnitude to the SProd condition in Experiment 1 (d=0.013, p=0.878) (**Figure 3, Table 4**). In the two spoken production experiments (Experiments 1 and 2a), the WProd condition elicited a stronger response than the NProd condition, suggesting that it was the most cognitively demanding production condition. Strong responses in the MD network during single-word production underscore the contributions of domain-general executive mechanisms to confrontation naming abilities—one of the most commonly used clinical language assessment tools (e.g., Kaplan et al., 1983).

4. Our whole-brain search for brain regions that respond more during sentence production than during sentence comprehension and word production did not reveal any regions that are replicably selective for sentence generation during language production. All of the regions that were recovered in this analysis displayed strong responses to word production, sentence comprehension, and/or visual event semantic conditions in addition to the sentence production condition (**Figures SI-3**, **SI-4**).

Discussion

We examined neural responses to word and sentence production in the language-selective network (Fedorenko et al., 2011), and across the brain, to illuminate the architecture of language generation. Across three experiments that employed a picture naming/description paradigm, we conceptually replicated prior studies (e.g., Menenti et al., 2011; Silbert et al., 2014) in observing robust responses in the language network to spoken sentence production. Further, we report three novel results: i) responses in the language network to sentence production are output-modality independent, with strong responses to both spoken and typed productions; ii) the language network responds to both lexical access and sentence generation demands; and iii) no brain region—within or outside of the language network—appears to be selective for sentence generation (relative to sentence comprehension and single-word production). Below, we contextualize these results in the current theoretical and empirical landscape of the field and discuss their implications.

Modality independence of language production

A key signature of the language network is input-modality independence during comprehension, as evidenced by similar responses across listening and reading (e.g., Fedorenko et al., 2010; Vagharchakian et al., 2012; Regev et al., 2013; Scott et al., 2017; Chen, Affourtit et al., 2021), as well as during visual processing in sign language comprehension (e.g., MacSweeney et al., 2008). Here, we show that the language network also exhibits modality-independent responses in production, across speaking and typing. This generalization across modalities demonstrates that the observed effects concern higher-level aspects of production (access of linguistic representations and utterance planning) rather than lower-level implementation parts of the production pipeline. As discussed in the Introduction, the network of areas that support articulation (e.g., Bohland & Guenther, 2006; Fedorenko & Thompson-Schill, 2014; Basilakos et al., 2018; see **Figure SI-2a** for a profile of regions in this network) is distinct from the higher-level language network examined here. The hand motor control areas associated with writing or typing linguistic utterances have been less extensively investigated, but are also distinct from the language network (e.g., Roux et al., 2009 Longcamp et al., 2014; Willet et al., 2021; **Figure SI-2b**).

Further, the fact that the language network responds during both interpretation and generation of linguistic utterances suggests that this network plausibly stores our language knowledge—mappings between forms and meanings—that are necessary for both comprehension (by evaluating the input relative to these representations) and production (by searching these representations for the right words/constructions). The fact that responses to sentence production are distributed across the language network aligns with growing evidence that this network constitutes a ‘natural kind’ in the mind and brain, working as an integrated system to solve comprehension and production (e.g., Mesulam, 1990; Blank et al. 2014; Fedorenko & Thompson-Schill, 2014; Braga et al., 2020).

In Experiment 2, where the same participants performed a spoken (Experiment 2a) and a typed (Experiment 2b) version of the production experiment, spoken sentence production elicited a stronger response than typed production. This effect seems surprising given the primacy of spoken language in ontogeny and phylogeny, and the intuitively greater effort associated with writing (but see next section for a note on the possible role of the lack of feedback during typing). Interestingly, we see a stronger response to typed than spoken production in the Multiple Demand (MD) network (**Figure 3**). This observation is notable because although the MD network responds to cognitive effort across domains (e.g., Duncan & Owen, 2000; Fedorenko et al., 2013; Hugdahl et al., 2015; Assem et al., 2020), *linguistic* demands have been shown to draw on the language-selective network (e.g., Diachek, Blank, Siegelman et al., 2020; Shain, Blank et al., 2020; Quillen et al., 2021; Wehbe et al., 2021; see Fedorenko & Shain, in press, for a review). The stronger response to typed production in the MD network therefore suggests that some demands related to processing language in a relatively-late-acquired modality may be supported by domain-general mechanisms, and points to a possible role of the MD network in learning to read and write.

Lack of selectivity for sentence generation relative to lexical access

We used a paradigm adapted from comprehension (e.g., Friederici et al., 2000; Humphries et al., 2006; Fedorenko et al., 2010) in an effort to separate lexical access and sentence generation demands. In comprehension, the question of whether different brain regions in the language network support single-word processing vs. combinatorial (syntactic/semantic) processing has long been controversial. Based on the critical review of the literature and several additional studies, Fedorenko et al. (2020) argued that no brain region within the language network is selective for combinatorial processing over the processing of single words (see also Chee et al., 1999; Keller et al., 2001; Röder et al., 2002; Bautista & Wilson, 2016; Blank et al., 2016; see

Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux et al., 2021 for converging evidence from relating human neural representations to those from artificial neural network models).

Here, we asked this question for language production. Single-word production requires a search through our linguistic knowledge store for the word that captures the intended meaning. Sentence production also requires accessing the relevant words, but it *additionally* requires assembling words and constructions into well-formed utterances, including ordering the words and selecting the right form of each word according to the intended meaning and the structure being built (e.g., selecting the plural form of a noun, or selecting the right tense for a verb). These additional cognitive operations plausibly lead to a stronger response during sentence production than single-word production (similar to what has been reported in comprehension; Fedorenko et al., 2010; Diachek, Blank, Siegelman et al., 2020), but critically, the language regions also show stronger responses to single-word production relative to low-level articulation. This pattern suggests that the same brain areas support lexical access and sentence construction during language production, mirroring the picture that has emerged in comprehension, and aligning with the general idea of memory as a computational resource in the brain (e.g., Hasson et al., 2015; Dasgupta & Gershman, 2020).

Some might argue that the sentences in our sentence production condition were too simple, and perhaps generating sentences with more complex structures would elicit syntax-selective responses. However, an architecture where different kinds of syntactic dependencies are supported by distinct mechanisms seems unlikely. Indeed, in language comprehension, the brain areas that are sensitive to simple two-word composition (e.g., Pallier et al., 2011; Shain et al., in prep.) are also engaged for the processing of non-local dependencies (e.g., Blank et al., 2016; Shain, Blank et al., 2020).

What about evidence from aphasia? Although some patients have been argued to exhibit a selective grammatical deficit ('agrammatism'; see deBleser, 1987 for a historical overview), the evidence is complex and does not point to the existence of syntax-selective machinery (e.g., Badecker & Caramazza, 1985; Berndt et al., 1996). *First*, patients with expressive agrammatism are a heterogeneous population: extreme variability has been reported both in production, including patients who produce complex structures (e.g., Stark & Dressler, 1990), and in comprehension (e.g., Goodglass & Menn, 1985; Howard, 1985; Berndt, 1987; Parisi, 1987), including patients who exhibit no difficulties in understanding complex structures (e.g., Miceli et al., 1983; Nespoulous et al., 1988). *Second*, neither agrammatic production nor comprehension have been consistently linked to damage to a particular brain region within the language network: damage to both frontal and temporal language areas and the white matter tracts connecting them have been shown to result in agrammatic production and/or comprehension (e.g., Kempler et al., 1991; Caplan et al., 1996; Dick et al., 2001; Wilson & Saygin, 2004; Mesulam et al., 2015), as was already discovered in the early days of research on agrammatism (deBleser, 1987). In line with this apparently distributed nature of syntactic processing, agrammatic performance has been shown to be inducible in neurotypical adults under cognitive load (e.g., Miyake et al., 1994; Blackwell & Bates, 1995). *Finally*, anomia (lexical retrieval difficulties) is ubiquitous in aphasia, including for patients with agrammatic production and/or comprehension (e.g., Goodglass & Geschwind, 1976; Blumstein, 1988; see Lu et al., 2021 for related evidence from cortical stimulation), suggesting that brain areas whose damage leads to agrammatism also contribute to lexical access (see also Bates & Goodman, 1997).

One experiment where we did not observe a response to lexical access demands (i.e., no stronger response during word production compared to nonword production, i.e., compared to articulation demands)—in the presence of robust sensitivity to sentence production demands—is Experiment 2b, where typed responses were used. We speculate that the lack of sensitivity to lexical access may have to do with the lack of visual feedback that typically accompanies written production (similar to auditory and proprioceptive feedback that accompanies speaking). The lack of such feedback may lead to shallower semantic engagement with the activated lexical representations in the context of unconnected words (cf. in the context of sentences, where words are assembled into complex semantic representations, which are more robust in memory; e.g., Potter, 2012).

Lack of production-selective syntactic encoding mechanisms

To test prior claims about the possible existence of production-selective syntactic encoding mechanisms (e.g., Bock, 1995), we searched for brain regions that would respond more strongly to sentence production than single-word production, and also than sentence comprehension. No such regions were found. (We acknowledge that some areas outside the boundaries of the language network, as defined here, may selectively contribute to *lexical access in production*, as has been suggested in some patient studies (e.g., Bi et al., 2011; Mesulam et al., 2013).) Instead, responses to sentence production were overall stronger than to sentence comprehension across the language network, suggesting that production is more costly. The latter is perhaps unsurprising: production trails comprehension in development and is more challenging for adult language learners (e.g., Jakobson 1941/1968).

Importantly, syntactic demands associated with sentence generation appear to be implemented within the language-selective network; the domain-general MD network, which supports demanding tasks across domains (Duncan, 2010, 2013), responds more strongly during the production of unrelated words than during sentence generation, similar to what has been observed for comprehension (e.g., Diachek, Blank, Siegelman et al., 2020; Fedorenko & Shain, in press).

In conclusion, we have shown that the language-selective network, which supports comprehension across modalities, also supports production during speaking and typing. Similar to the strong integration between word meanings and combinatorial processing that has been observed for comprehension, we found that the language areas support both lexical access and sentence generation during production. These results support the idea that this network stores integrated linguistic knowledge, from phonotactic regularities, to morphological schemas, to words, to constructions. Finally, contra prior hypotheses, we did not find evidence of brain areas selective for syntactic encoding during production relative to comprehension; instead, sentence production appears to pose a higher cost to the language network than sentence comprehension.

Figures

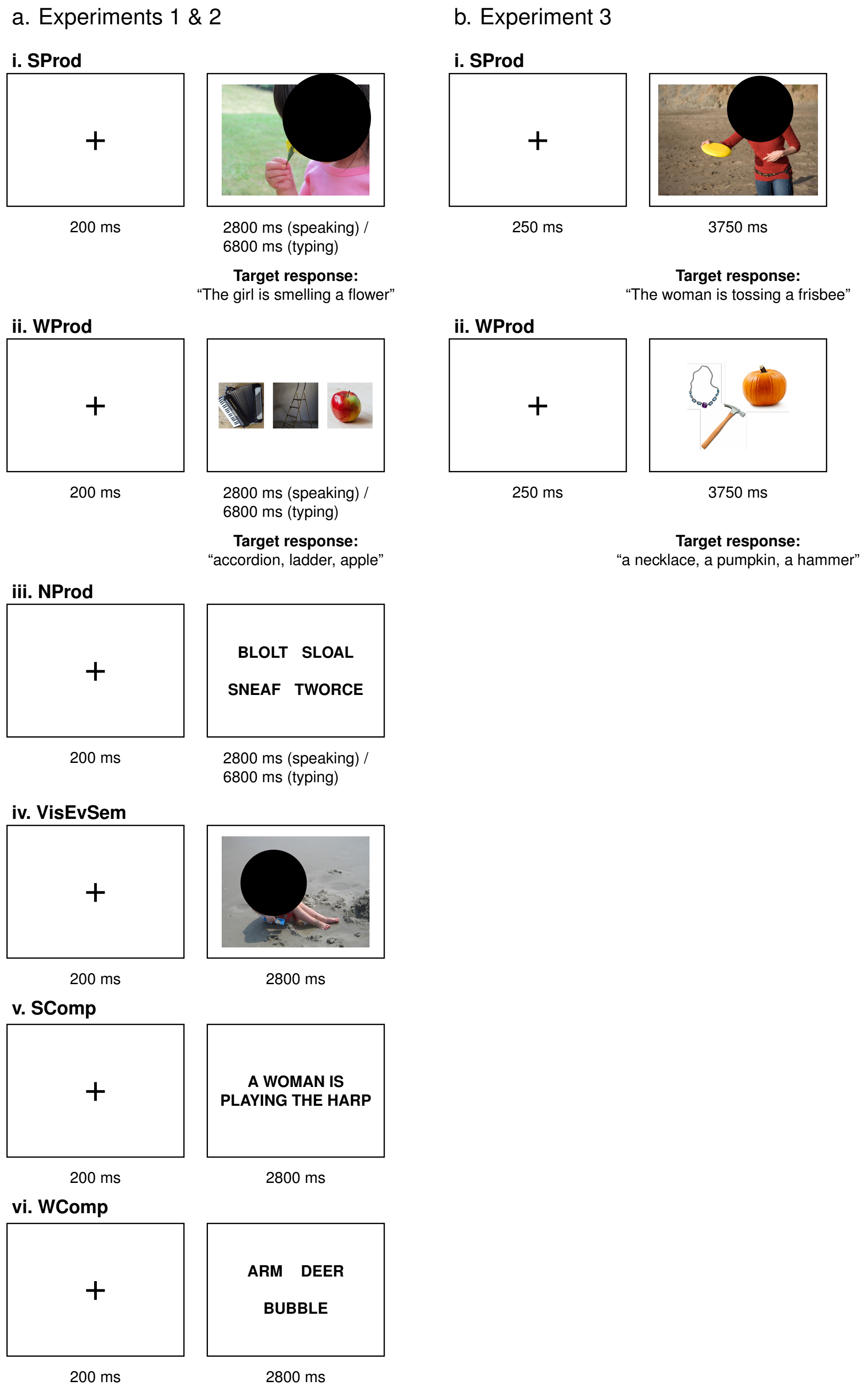
a. Experiments 1 & 2
i. SProd
+
200 ms
2800 ms (speaking) / 6800 ms (typing)
Target response:
"The girl is smelling a flower"
ii. WProd
+
200 ms
2800 ms (speaking) / 6800 ms (typing)
Target response:
"accordion, ladder, apple"
iii. NProd
+
BLOLT SLOAL
SNEAF TWORCE
200 ms
2800 ms (speaking) / 6800 ms (typing)
iv. VisEvSem
+
200 ms
2800 ms
v. SComp
+
A WOMAN IS PLAYING THE HARP
200 ms
2800 ms
vi. WComp
+
ARM DEER
BUBBLE
200 ms
2800 ms
b. Experiment 3
i. SProd
+
250 ms
3750 ms
Target response:
"The woman is tossing a frisbee"
ii. WProd
+
250 ms
3750 ms
Target response:
"a necklace, a pumpkin, a hammer"

Figure 1. Sample trials for each condition of Experiments 1-3. NOTE: faces have been redacted from event photographs (black circles); for the original images, see https://github.com/jennhu/LanguageProduction/. **a.** In Experiments 1 and 2, participants performed two production tasks: producing descriptions of events depicted in naturalistic photographs (SProd), and producing names of unrelated, isolated objects depicted in separate photographs (WProd). In two control conditions, participants spoke or typed monosyllabic nonwords (NProd), and indicated whether events depicted in photographs took place indoors or outdoors (VisEvSem). Finally, participants performed two comprehension tasks: silently reading sentences (SComp) and word lists (WComp), mirroring the structure and content of target responses from the production trials (see Methods for details). Trial durations were increased from 2800ms to 6800ms for conditions requiring typed (cf. spoken) output in Experiment 2b. **b.** In Experiment 3, participants performed the sentence and word production conditions (SProd, WProd) with a different set of materials.

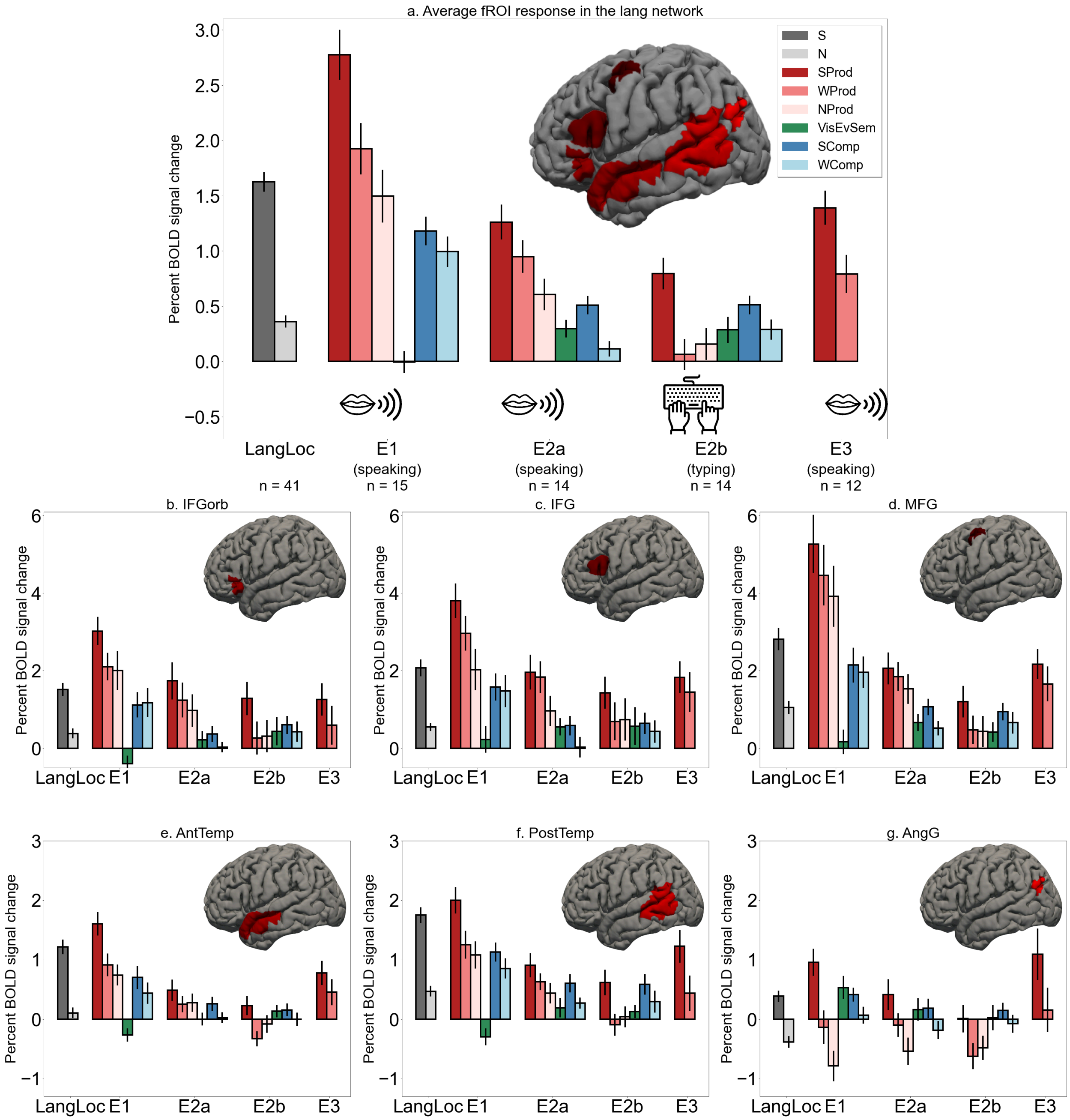

Figure 2. Responses in the language network. Responses in the language network to the language localizer conditions (dark grey=sentences (S), light grey=nonwords (N); the responses are pooled across all participants) and the conditions of the critical experiments (red shades=production conditions (from darker to lighter: sentence production (SProd), word-list production (WProd), and nonword-list production (NProd); green=visual event semantic

condition (VisEvSem); blue shades=comprehension conditions (dark=sentence comprehension (SComp), light=word-list comprehension (WComp)) in Experiments 1-3. (Notice that the overall magnitude of responses in Experiment 1 is higher than in Experiments 2 and 3. Such differences across groups of participants are not uncommon, and plausibly attributable to trait/state differences (e.g., Hajnal et al., 1994; He et al., 2010; Schölvinck et al., 2010; Poldrack, 2011; Wong et al., 2013; J. Power et al., 2015; Chang et al., 2016; Erdogan et al., 2016).) The top panel shows the responses averaging across the six regions of the language network. On the brain inset, we show the parcels that were used to define the individual language functional ROIs (any individual fROI is 10% of the parcel, as described in the Methods). The bottom panels show the responses for each of the six regions of the language network. Error bars represent standard errors of the mean over participants.

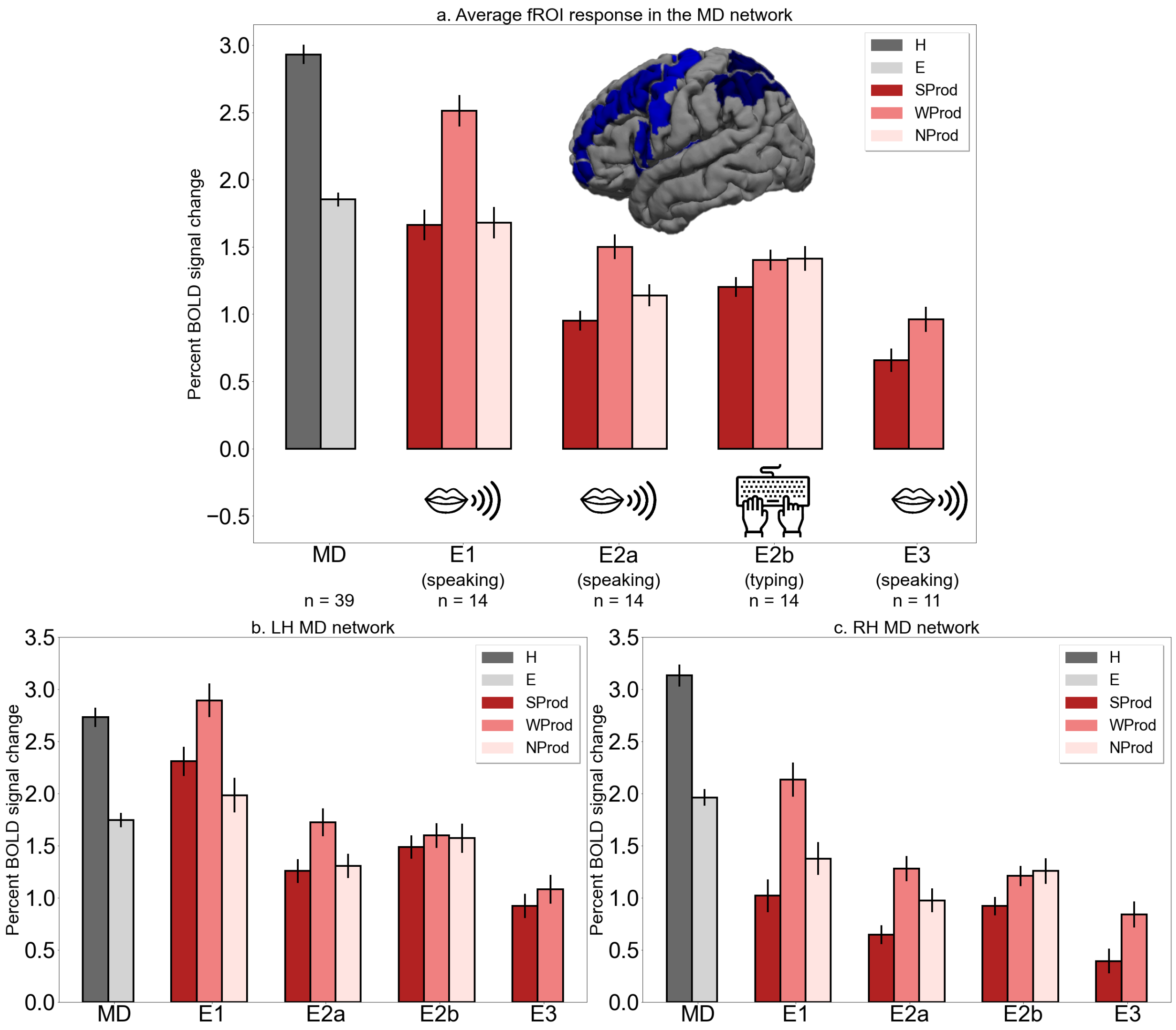

Figure 3. Responses in the MD network. Responses in the multiple demand (MD) network to the MD localizer conditions (dark grey=hard spatial working memory (H), light grey=easy spatial working memory (E); the responses are pooled across all participants) and the production conditions of the critical experiments (from darker to lighter red shades: sentence production (SProd), word-list production (WProd), and nonword-list production (NProd)) in Experiments 1-3. The top panel shows the responses averaging across the 20 regions of the MD network. On the brain inset, we show the parcels that were used to define the individual MD functional ROIs (any individual fROI is 10% of the parcel, as described in the Methods). The bottom panels show the

responses averaged over the 10 regions in each of the left and right hemispheres of the MD network. Error bars represent standard errors of the mean over participants.

Tables

	Experiment 1			Experiment 2a			Experiment 3	
fROI	SProd vs. fixation	SProd vs. N	SProd vs. VisEvSem	SProd vs. fixation	SProd vs. N	SProd vs. VisEvSem	SProd vs. fixation	SProd vs. N
IFGorb	$d = 4.042$ $p < 0.001$	$d = 3.069$ $p < 0.001$	$d = 5.665$ $p < 0.001$	$d = 1.835$ $p = 0.004$	$d = 1.639$ $p = 0.033$	$d = 1.122$ $p = 0.016$	$d = 1.656$ $p = 0.012$	$d = 2.300$ $p = 0.003$
IFG	$d = 4.183$ $p < 0.001$	$d = 3.939$ $p < 0.001$	$d = 4.292$ $p < 0.001$	$d = 2.213$ $p = 0.001$	$d = 2.082$ $p = 0.014$	$d = 1.754$ $p = 0.016$	$d = 2.449$ $p = 0.001$	$d = 2.452$ $p = 0.003$
MFG	$d = 3.409$ $p < 0.001$	$d = 2.777$ $p < 0.001$	$d = 3.892$ $p < 0.001$	$d = 2.545$ $p < 0.001$	$d = 1.530$ $p = 0.033$	$d = 1.184$ $p = 0.016$	$d = 3.115$ $p < 0.001$	$d = 3.300$ $p = 0.001$
AntTemp	$d = 3.928$ $p < 0.001$	$d = 3.559$ $p < 0.001$	$d = 4.967$ $p < 0.001$	$d = 1.374$ $p = 0.019$	$d = 1.161$ $p = 0.068$	$d = 0.904$ $p = 0.035$	$d = 2.105$ $p = 0.003$	$d = 2.519$ $p = 0.003$
PostTemp	$d = 4.391$ $p < 0.001$	$d = 3.561$ $p < 0.001$	$d = 4.785$ $p < 0.001$	$d = 2.251$ $p = 0.001$	$d = 1.225$ $p = 0.068$	$d = 1.072$ $p = 0.017$	$d = 2.434$ $p = 0.001$	$d = 2.410$ $p = 0.003$
AngG	$d = 2.058$ $p < 0.001$	$d = 3.319$ $p < 0.001$	$d = 1.688$ $p = 0.007$	$d = 0.630$ $p = 0.128$	$d = 0.928$ $p = 0.118$	$d = 0.702$ $p = 0.228$	$d = 1.393$ $p = 0.025$	$d = 1.575$ $p = 0.003$
Language network	$d = 2.241$ $p < 0.001$	$d = 1.832$ $p < 0.001$	$d = 2.081$ $p < 0.001$	$d = 1.527$ $p < 0.001$	$d = 1.054$ $p < 0.001$	$d = 1.037$ $p < 0.001$	$d = 1.731$ $p < 0.001$	$d = 1.587$ $p < 0.001$

Table 1. Responses in the language network to spoken sentence production. Effect sizes (Cohen's *d*) and estimated *p*-values for the effect of the spoken SProd condition (relative to three baselines) in linear mixed-effects regression models in Experiments 1, 2a, and 3 (see Analyses, Q1). For each experiment that included the spoken sentence production condition (Experiments 1, 2a, and 3), models were fit to perform pairwise comparisons of sentence production (SProd) vs. fixation, and sentence production vs. nonword comprehension (N; from the language localizer). In Experiments 1 and 2a, we additionally compared sentence production vs. the visual event semantics condition (VisEvSem). The results are shown at the level of individual functional ROIs in the language network (first six rows; FDR corrected), as well as averaged across the language network (last row). Green cells highlight significance at $p<0.05$ in the predicted direction.

fROI	**Experiment 2b**			
	SProd (typed) vs. fixation	SProd (typed) vs. N	SProd (typed) vs. VisEvSem	SProd (typed) vs. SProd
IFGorb	d = 1.507 p = 0.018	d = 1.263 p = 0.121	d = 1.603 p = 0.025	d = 0.501 p = 0.383
IFG	d = 1.771 p = 0.018	d = 1.283 p = 0.121	d = 1.650 p = 0.025	d = 0.614 p = 0.346
MFG	d = 1.494 p = 0.018	d = 0.361 p = 0.631	d = 1.492 p = 0.028	d = 0.580 p = 0.333
AntTemp	d = 0.561 p = 0.198	d = 0.274 p = 0.631	d = 0.326 p = 0.680	d = 0.813 p = 0.333
PostTemp	d = 1.443 p = 0.018	d = 0.488 p = 0.631	d = 1.764 p = 0.025	d = 0.707 p = 0.337
AngG	d = 0.015 p = 0.969	d = 0.006 p = 0.992	d = −0.077 p = 0.892	d = 1.216 p = 0.283
Language network	d = 1.027 p < 0.001	d = 0.498 p = 0.003	d = 0.576 p < 0.001	d = 0.459 p = 0.006

Table 2. Responses in the language network to typed sentence production, and a comparison of typed vs. spoken production. Effect sizes (Cohen's *d*) and estimated *p*-values for the effect of the typed SProd condition (relative to three baseline and to the spoken SProd condition) in linear mixed-effects regression models (see Analyses, Q2) in Experiment 2. Models were fit to perform pairwise comparisons of typed sentence production (SProd) vs. fixation, typed sentence production vs. nonword comprehension (N; from the language network localizer), typed sentence production vs. the visual event semantics condition (VisEvSem), and typed sentence production vs. spoken sentence production (from Experiment 2a). The results are shown at the level of individual functional ROIs in the language network (first six rows; FDR corrected), as well as averaged across the language network (last row). Green cells highlight significance at p<0.05 in the predicted direction. Blue cells indicate significance at p<0.05 when no direction was predicted.

	Experiment 1		**Experiment 2a**		**Experiment 2b**		**Experiment 3**
fROI	WProd vs. NProd	SProd vs. WProd	WProd vs. NProd	SProd vs. WProd	WProd (typed) vs. NProd (typed)	SProd (typed) vs. WProd (typed)	SProd vs. WProd
IFGorb	$d = 0.184$ $p = 0.736$	$d = 2.503$ $p < 0.001$	$d = 0.659$ $p = 0.308$	$d = 1.321$ $p = 0.081$	$d = -0.150$ $p = 0.825$	$d = 2.966$ $p < 0.001$	$d = 2.727$ $p = 0.002$
IFG	$d = 1.506$ $p = 0.082$	$d = 1.867$ $p = 0.004$	$d = 1.572$ $p = 0.085$	$d = 0.290$ $p = 0.610$	$d = -0.125$ $p = 0.825$	$d = 1.570$ $p = 0.014$	$d = 1.003$ $p = 0.124$
MFG	$d = 1.064$ $p = 0.133$	$d = 1.628$ $p = 0.009$	$d = 0.717$ $p = 0.308$	$d = 0.497$ $p = 0.464$	$d = 0.148$ $p = 0.825$	$d = 1.942$ $p = 0.005$	$d = 1.158$ $p = 0.097$
AntTemp	$d = 0.620$ $p = 0.318$	$d = 2.419$ $p < 0.001$	$d = -0.122$ $p = 0.829$	$d = 1.264$ $p = 0.081$	$d = -1.881$ $p = 0.029$	$d = 2.988$ $p < 0.001$	$d = 2.251$ $p = 0.005$
PostTemp	$d = 0.771$ $p = 0.257$	$d = 3.795$ $p < 0.001$	$d = 0.672$ $p = 0.308$	$d = 1.044$ $p = 0.124$	$d = -0.831$ $p = 0.378$	$d = 3.693$ $p < 0.001$	$d = 2.816$ $p = 0.002$
AngG	$d = 1.273$ $p = 0.096$	$d = 3.529$ $p < 0.001$	$d = 1.136$ $p = 0.184$	$d = 1.943$ $p = 0.023$	$d = -0.769$ $p = 0.378$	$d = 2.718$ $p < 0.001$	$d = 4.396$ $p < 0.001$
Language network	$d = 0.333$ $p = 0.037$	$d = 0.668$ $p < 0.001$	$d = 0.376$ $p = 0.023$	$d = 0.356$ $p = 0.032$	$d = -0.105$ $p = 0.524$	$d = 0.830$ $p < 0.001$	$d = 0.609$ $p < 0.001$

Table 3. Responses in the language network to lexical access and sentence generation.

Effect sizes (Cohen's *d*) and estimated *p*-values for the effect of WProd condition (relative to the NProd condition) and the effect of SProd condition (relative to the WProd condition) in linear mixed-effects regression models in Experiments 1-3 (see Analyses, Q3). For each of Experiments 1, 2a, and 2b, models were fit to perform pairwise comparisons of word list production (WProd) vs. nonword production (NProd)—a contrast that targets lexical access, and sentence production (SProd) vs. word list production (WProd)—a contrast that targets sentence generation. For Experiment 3, which lacked the NProd condition, we only compared sentence production vs. word list production. The results are shown at the level of individual functional ROIs in the language network (first six rows; FDR corrected), as well as averaged across the language network (last row). Green cells highlight significance at p<0.05 in the predicted direction. Red cells highlight significance at p<0.05 in the non-predicted direction.

fROI	**Experiment 1**		**Experiment 2a**		**Experiment 2b**		**Experiment 3**
	SProd vs. WProd	SProd vs. NProd	SProd vs. WProd	SProd vs. NProd	SProd vs. WProd	SProd vs. NProd	SProd vs. WProd
LH PostParietal	$d = -3.491$ $p < 0.001$	$d = 0.871$ $p = 0.200$	$d = -1.873$ $p = 0.059$	$d = -0.043$ $p = 0.989$	$d = -0.962$ $p = 0.354$	$d = -1.837$ $p = 0.039$	$d = -1.034$ $p = 0.190$
LH midParietal	$d = -1.731$ $p = 0.014$	$d = -0.066$ $p = 0.907$	$d = -1.137$ $p = 0.076$	$d = -1.053$ $p = 0.267$	$d = -0.530$ $p = 0.594$	$d = -1.724$ $p = 0.039$	$d = 0.912$ $p = 0.240$
LH antParietal	$d = -1.266$ $p = 0.053$	$d = -1.793$ $p = 0.033$	$d = -1.194$ $p = 0.073$	$d = -1.851$ $p = 0.060$	$d = -0.690$ $p = 0.483$	$d = -1.706$ $p = 0.039$	$d = 1.413$ $p = 0.114$
LH supFrontal	$d = -1.156$ $p = 0.064$	$d = 0.964$ $p = 0.200$	$d = -0.767$ $p = 0.200$	$d = 0.072$ $p = 0.989$	$d = -0.267$ $p = 0.769$	$d = -0.661$ $p = 0.364$	$d = -1.132$ $p = 0.160$
LH Precentral A PrecG	$d = -1.392$ $p = 0.037$	$d = -0.968$ $p = 0.200$	$d = -1.400$ $p = 0.064$	$d = -0.721$ $p = 0.443$	$d = -0.317$ $p = 0.769$	$d = -1.093$ $p = 0.128$	$d = -1.398$ $p = 0.114$
LH Precentral B IFGop	$d = -1.232$ $p = 0.055$	$d = 0.601$ $p = 0.373$	$d = -0.927$ $p = 0.132$	$d = 0.171$ $p = 0.898$	$d = -0.099$ $p = 0.906$	$d = 0.181$ $p = 0.750$	$d = -1.132$ $p = 0.160$
LH midFrontal	$d = -1.221$ $p = 0.055$	$d = 1.895$ $p = 0.031$	$d = -1.473$ $p = 0.064$	$d = 0.192$ $p = 0.898$	$d = -0.444$ $p = 0.673$	$d = 0.364$ $p = 0.616$	$d = -0.521$ $p = 0.477$
LH midFrontalOrb	$d = -1.036$ $p = 0.089$	$d = 2.235$ $p = 0.019$	$d = -1.174$ $p = 0.073$	$d = 0.184$ $p = 0.898$	$d = 0.040$ $p = 0.943$	$d = 0.558$ $p = 0.416$	$d = -0.828$ $p = 0.275$
LH insula	$d = -1.601$ $p = 0.020$	$d = 0.908$ $p = 0.200$	$d = -1.092$ $p = 0.083$	$d = 0.723$ $p = 0.443$	$d = 0.681$ $p = 0.483$	$d = 1.627$ $p = 0.039$	$d = -3.099$ $p = 0.006$
LH medialFrontal	$d = -0.990$ $p = 0.098$	$d = 2.157$ $p = 0.019$	$d = -0.652$ $p = 0.261$	$d = 0.678$ $p = 0.443$	$d = 0.351$ $p = 0.769$	$d = 1.638$ $p = 0.039$	$d = -1.290$ $p = 0.134$
RH PostParietal	$d = -3.255$ $p < 0.001$	$d = 0.922$ $p = 0.200$	$d = -1.622$ $p = 0.059$	$d = 0.621$ $p = 0.473$	$d = -1.432$ $p = 0.223$	$d = -1.376$ $p = 0.055$	$d = -3.804$ $p = 0.003$
RH midParietal	$d = -2.518$ $p = 0.001$	$d = -1.318$ $p = 0.096$	$d = -1.363$ $p = 0.064$	$d = -1.637$ $p = 0.063$	$d = -1.233$ $p = 0.223$	$d = -1.802$ $p = 0.039$	$d = -0.098$ $p = 0.879$
RH antParietal	$d = -2.543$ $p = 0.001$	$d = -1.584$ $p = 0.054$	$d = -1.364$ $p = 0.064$	$d = -1.818$ $p = 0.060$	$d = -0.946$ $p = 0.354$	$d = -1.468$ $p = 0.047$	$d = -0.695$ $p = 0.350$
RH supFrontal	$d = -2.369$ $p = 0.002$	$d = -0.876$ $p = 0.200$	$d = -1.662$ $p = 0.059$	$d = -0.830$ $p = 0.443$	$d = -1.250$ $p = 0.223$	$d = -1.561$ $p = 0.042$	$d = -2.722$ $p = 0.008$
RH Precentral A PrecG	$d = -2.397$ $p = 0.002$	$d = -0.997$ $p = 0.200$	$d = -1.283$ $p = 0.073$	$d = -0.683$ $p = 0.443$	$d = -1.470$ $p = 0.223$	$d = -1.455$ $p = 0.047$	$d = -1.262$ $p = 0.134$
RH Precentral B IFGop	$d = -1.720$ $p = 0.014$	$d = -0.298$ $p = 0.666$	$d = -1.181$ $p = 0.073$	$d = -0.459$ $p = 0.650$	$d = -0.614$ $p = 0.524$	$d = -0.949$ $p = 0.185$	$d = -0.452$ $p = 0.517$
RH midFrontal	$d = -3.083$ $p < 0.001$	$d = -0.612$ $p = 0.373$	$d = -1.633$ $p = 0.059$	$d = -1.152$ $p = 0.233$	$d = -0.913$ $p = 0.354$	$d = -0.563$ $p = 0.416$	$d = -2.101$ $p = 0.022$
RH midFrontalOrb	$d = -2.596$ $p = 0.001$	$d = -1.372$ $p = 0.093$	$d = -1.386$ $p = 0.064$	$d = -1.603$ $p = 0.063$	$d = -0.720$ $p = 0.483$	$d = -0.336$ $p = 0.617$	$d = -2.201$ $p = 0.020$
RH insula	$d = -3.602$ $p < 0.001$	$d = -0.397$ $p = 0.572$	$d = -1.193$ $p = 0.073$	$d = -0.326$ $p = 0.810$	$d = -0.242$ $p = 0.769$	$d = 0.303$ $p = 0.625$	$d = -2.960$ $p = 0.006$
RH medialFrontal	$d = -2.505$ $p = 0.001$	$d = -0.102$ $p = 0.902$	$d = -1.171$ $p = 0.073$	$d = -0.002$ $p = 0.997$	$d = 0.224$ $p = 0.769$	$d = 0.735$ $p = 0.320$	$d = -2.384$ $p = 0.015$
MD network	$d = -0.692$ $p < 0.001$	$d = -0.013$ $p = 0.878$	$d = -0.612$ $p < 0.001$	$d = -0.212$ $p = 0.015$	$d = -0.225$ $p = 0.010$	$d = -0.228$ $p = 0.009$	$d = -0.353$ $p < 0.001$

Table 4. Responses in the MD network to spoken production conditions. Effect sizes (Cohen's *d*) and estimated *p*-values for effect of the spoken SProd condition (relative to two baselines) in mixed-effects regression models in Experiments 1-3 (see Analyses, Q4). For each experiment, models were fit to perform pairwise comparisons of spoken sentence production

(SProd) vs. word-list production (WProd). In Experiments 1, 2a, and 2b, we additionally compared sentence production vs. nonword-list production (NProd). The results are shown at the level of individual functional ROIs in the multiple demand (MD) network (first 20 rows; FDR corrected), as well as averaged across the MD network (last row). Green cells highlight significance at $p<0.05$ in the predicted direction. Red cells highlight significance at $p<0.05$ in the non-predicted direction.

Supplementary Information

Experiment	# subjects	Trial Structure	# trials & blocks	Task	Total duration
Experiments 1, 2, and 3	15, 14, 6	100ms trial-initial fixation; 12 words/nonwords presented for 450ms each; 400ms button press; 100ms trial-final fixation	3 trials per block, 8 blocks per condition, 16 blocks total	Button press	5 minutes, 58 seconds
Experiment 3	4	300ms trial-initial fixation; 12 words/nonwords presented for 350 ms each; 1000 ms probe; 500 ms trial-final fixation	3 trials per block, 8 blocks per condition, 16 blocks total	Memory probe	6 minutes, 18 seconds
Experiment 3	2	300ms trial-initial fixation; 12 words/nonwords presented for 350 ms each; 1000 ms probe; 500 ms trial-final fixation	3 trials per block, 6 blocks per condition, 18 blocks total	Memory probe	6 minutes, 36 seconds

Table SI-1. Language localizer details for all experiments.

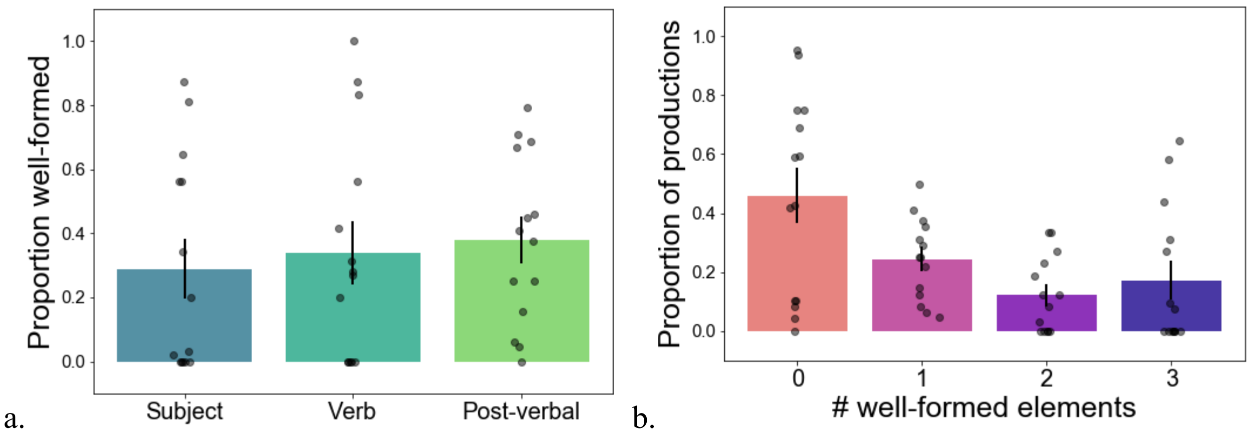

Figure SI-1. Details on the sentence productions in Experiment 2b. a. Mean proportions of syntactically well-formed (cf. 'newspaper headline' syntax) subject phrases, verb phrases, and post-verbal elements. For the subjects and post-verbal elements, we counted a phrase as well-formed if it included a determiner, and for verb phrases, we counted a phrase as well-formed if it used a tense form. Here, and in b, error bars represent standard errors of the mean over participants; individual dots represent participants. **b.** Mean proportion of productions with 0, 1, 2, or 3 well-formed elements.

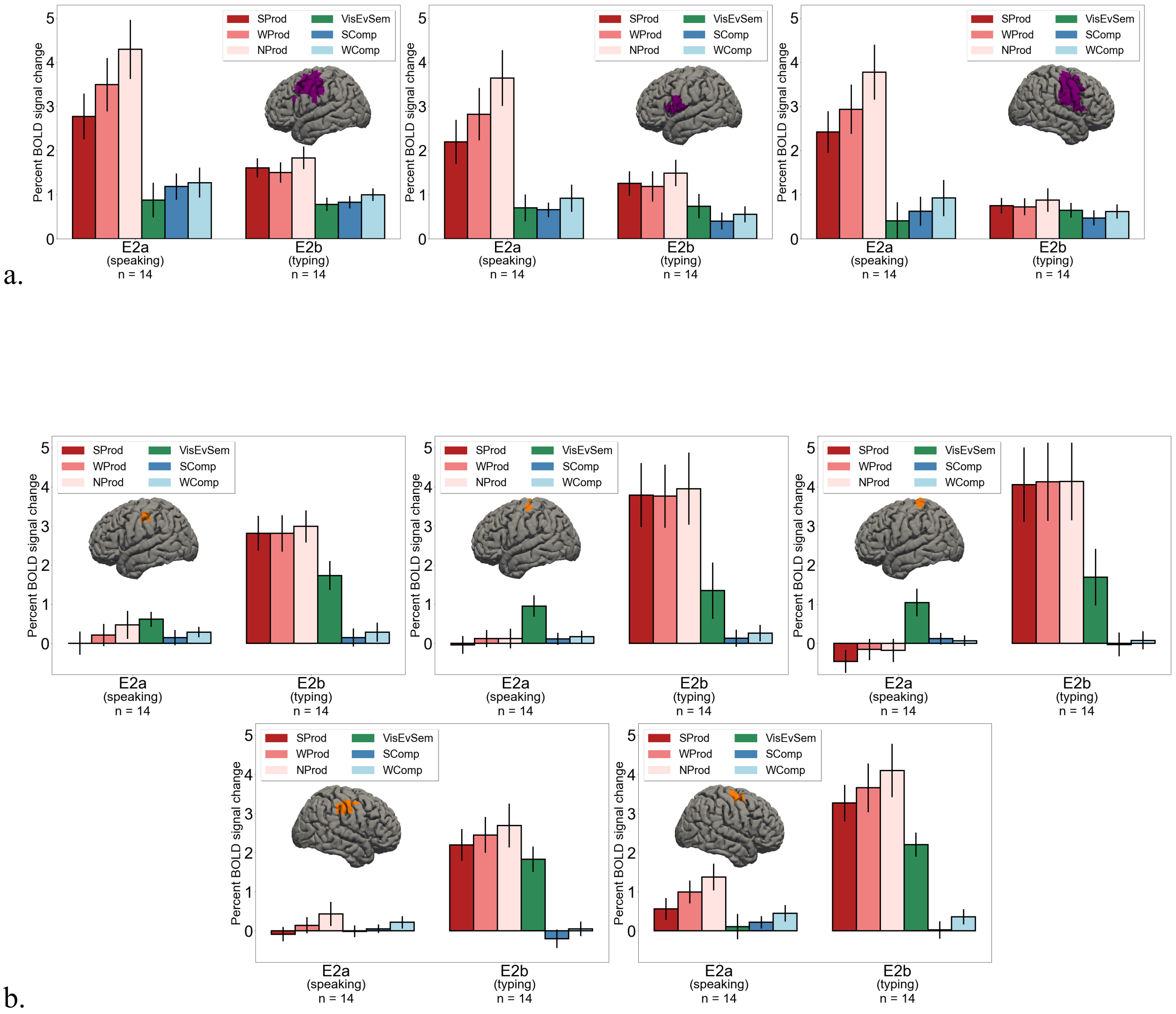

Figure SI-2. Response profiles of select articulation-selective (a) and typing-selective (b) areas. We performed a whole-brain group-constrained subject-specific (GSS) analysis (Fedorenko et al., 2010) for the nonword-list production (NProd) > Fixation contrast, using data from Experiment 2a (spoken production) to search for articulation-responsive areas and data from Experiment 2b (typed production) to search for typing-responsive areas. In particular, for each participant, we thresholded the whole-brain map for the relevant (spoken or typed) NProd>Fixation contrast at p<0.001 (uncorrected level). Each participant's map was then binarized, with 1s corresponding to voxels that show a reliable effect, and 0s otherwise. These individual maps were then overlaid in the common space, and watershed parcellation was performed, as described in Fedorenko et al. (2010), to search for areas that would contain supra-threshold voxels in at least half of the participants. The resulting regions (parcels) were then used to define the individual fROIs using the same contrast (NProd(spoken/typed)>Fix) by selecting the top 10% of voxels based on the *t*-values for the relevant contrast. In this way, a fROI was defined in every participant. Finally, we estimated the responses to the critical conditions (SProd, WProd, NProd, VisEvSem, SComp, and WComp) in these individually defined fROIs (to estimate the response to the NProd condition, across-runs cross-validation was used to ensure independence; as described in Methods for estimating the responses to the language and MD localizer conditions in the language and MD fROIs, respectively). **a.** Responses to the critical conditions in Experiments 2a and 2b for three sample articulation-responsive functional ROIs (parcels used to define these are shown as brain insets). **b.** Responses to the critical conditions in Experiments 2a and 2b, for five sample typing-responsive functional ROIs. As can be seen from the response profiles, these areas show the expected selectivity for spoken (figures in a) and typed (figures in b) language production. Unlike the language fROIs examined in the main text, these regions respond similarly strongly to language production regardless of the content (sentences, word lists, or nonword lists), as expected given the lower-level articulatory/hand-motor areas. (Note that the relatively strong responses to the VisEvSem condition are likely due to the fact that participants were using their fingers to perform the button press.)

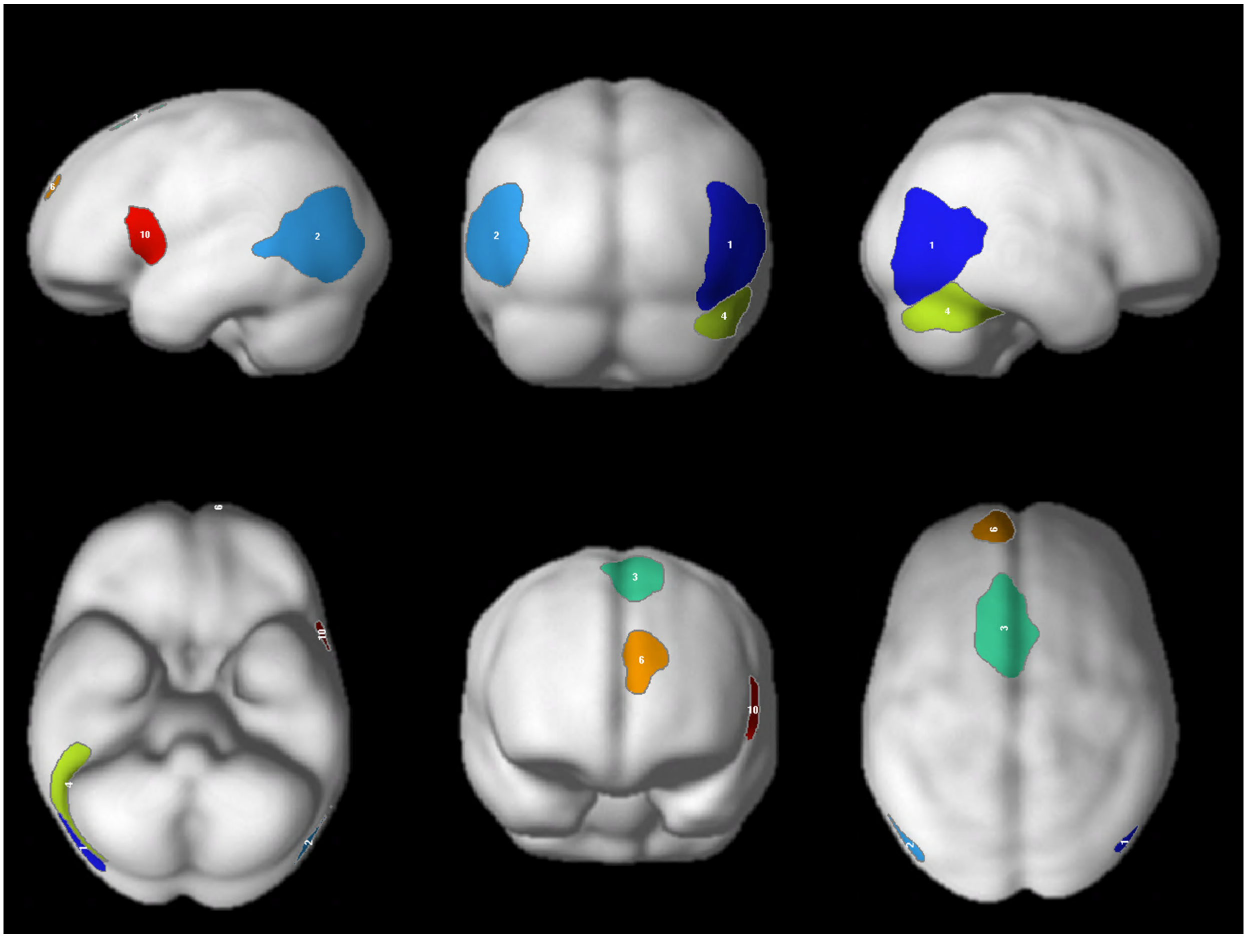

Figure SI-3. The locations of candidate sentence-production-selective regions identified by a whole-brain group-constrained subject-specific (GSS) analysis (Fedorenko et al., 2010) performed on the data from Experiments 1 and 2 on the conjunction of two contrasts (see Q4 in Methods): i) sentence production > sentence comprehension (SProd>SComp; thresholded at p<0.001 uncorrected level), and ii) sentence production > word-list production (SProd>WProd; thresholded at p <0.05 uncorrected level; the thresholds were selected based on visual inspection of the whole-brain activation maps for the relevant contrasts; note that the use of liberal thresholds is not problematic here given that all the results are assessed with cross-validation across independent data folds). For each participant, we thresholded and binarized the whole-brain maps for the relevant contrasts. These individual maps were then overlaid in the common space, and watershed parcellation was performed, as described in Fedorenko et al. (2010), to search for areas that would contain supra-threshold voxels for both contrasts in at least half of the participants. The resulting regions (parcels) were then used to define the individual fROIs using the same conjunction of contrasts by selecting the top 10% of voxels based on the *t*-values for each relevant contrast and taking the intersection of those voxels. Six areas were recovered that contained supra-threshold voxels in more than half of the participants (range: 0.55-0.93) and that showed replicable (as assessed using across-runs cross-validation) SProd>SComp and SProd>WProd effects. These areas' functional response profiles were then examined (see **Figure SI-4**).

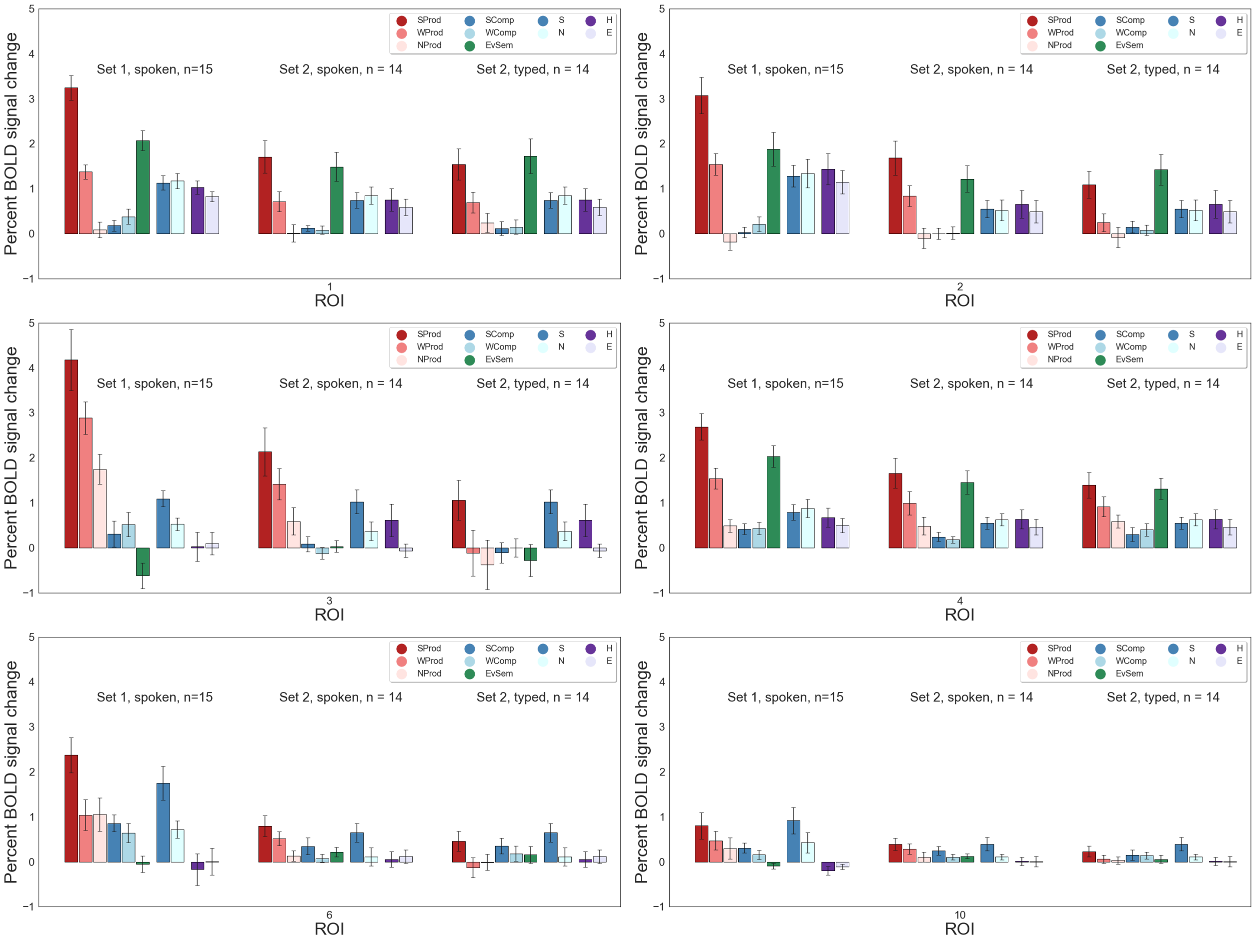

Figure SI-4. Response profiles of the candidate sentence-production-selective areas.

Responses in the candidate sentence-production-selective areas to the conditions of Experiments 1 and 2. The ROI numbers correspond to the parcel numbers in **Figure SI-3**. As can be seen, none of these candidate areas showed selectivity for sentence production based on their full response profile.

References

Assem, M., Glasser, M. F., Van Essen, D. C., & Duncan, J. (2020). A Domain-General Cognitive Core Defined in Multimodally Parcellated Human Cortex. *Cerebral Cortex*, *30*(8), 4361–4380. https://doi.org/10.1093/cercor/bhaa023

Badecker, W., & Caramazza, A. (1985). On considerations of method and theory governing the use of clinical categories in neurolinguistics and cognitive neuropsychology: The case against agrammatism. *Cognition*, *20*(2), 97–125. https://doi.org/10.1016/0010-0277(85)90049-6

Basilakos, A., Smith, K. G., Fillmore, P., Fridriksson, J., & Fedorenko, E. (2018). Functional Characterization of the Human Speech Articulation Network. *Cerebral Cortex*, *28*(5), 1816–1830. https://doi.org/10.1093/cercor/bhx100

Bates, E., & Goodman, J. C. (1997). On the Inseparability of Grammar and the Lexicon: Evidence from Acquisition, Aphasia and Real-time Processing. *Language and Cognitive Processes*, *12*(5–6), 507–584. https://doi.org/10.1080/016909697386628

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bautista, A., & Wilson, S. M. (2016). Neural responses to grammatically and lexically degraded speech. *Language, Cognition and Neuroscience*, *31*(4), 567–574. https://doi.org/10.1080/23273798.2015.1123281

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

Berndt, R. S. (1987). Symptom co-occurrence and dissociation in the interpretation of agrammatism. In M. Coltheart, G. Sartori, & R. Job (Eds.), *The Cognitive Neuropsychology of Language*. Erlbaum.

Berndt, R. S., Mitchum, C. C., & Haendiges, A. N. (1996). Comprehension of reversible sentences in "agrammatism": A meta-analysis. *Cognition*, *58*(3), 289–308. https://doi.org/10.1016/0010-0277(95)00682-6

Bi, Y., Wei, T., Wu, C., Han, Z., Jiang, T., & Caramazza, A. (2011). The role of the left anterior temporal lobe in language processing revisited: Evidence from an individual with ATL resection. *Cortex*, *47*(5), 575–587. https://doi.org/10.1016/j.cortex.2009.12.002

Blackwell, A., & Bates, E. (1995). Inducing Agrammatic Profiles in Normals: Evidence for the Selective Vulnerability of Morphology under Cognitive Resource Limitation. *Journal of Cognitive Neuroscience*, *7*(2), 228–257. https://doi.org/10.1162/jocn.1995.7.2.228

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

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *NeuroImage*, *127*, 307–323. https://doi.org/10.1016/j.neuroimage.2015.11.069

Blumstein, S. E. (1988). Neurolinguistics: An overview of language–brain relations in aphasia. In F. J. Newmeyer (Ed.), *Linguistics: The Cambridge Survey: Volume 3: Language: Psychological and Biological Aspects* (Vol. 3, pp. 210–236). Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9780511621062.009

Bock, K. (1995). Sentence Production: From Mind to Mouth. In J. L. Miller & P. D. Eimas (Eds.), *Speech, Language, and Communication* (2nd ed., pp. 181–216). Academic Press. https://www.sciencedirect.com/science/article/pii/B978012497770950008X

Bock, K. (1996). Language production: Methods and methodologies. *Psychonomic Bulletin & Review*, *3*(4), 395–421. https://doi.org/10.3758/BF03214545

Bohland, J. W., & Guenther, F. H. (2006). An fMRI investigation of syllable sequence production. *NeuroImage*, *32*(2), 821–841. https://doi.org/10.1016/j.neuroimage.2006.04.173

Borovsky, A., Saygin, A. P., Bates, E., & Dronkers, N. (2007). Lesion correlates of conversational speech production deficits. *Neuropsychologia*, *45*(11), 2525–2533. https://doi.org/10.1016/j.neuropsychologia.2007.03.023

Bouchard, K. E., Mesgarani, N., Johnson, K., & Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature*, *495*(7441), 327–332. https://doi.org/10.1038/nature11911

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

Brener, R. (1940). An experimental investigation of memory span. *Journal of Experimental Psychology*, *26*(5), 467–482. https://doi.org/10.1037/h0061096

Broca, P. (1861). Remarks on the Seat of the Faculty of Articulated Language, Following an Observation of Aphemia (Loss of Speech). *Bulletin de La Société Anatomique*, *6*, 330–357.

Caplan, D., Hildebrandt, N., & Makris, N. (1996). Location of lesions in stroke patients with deficits in syntactic processing in sentence comprehension. *Brain*, *119*(3), 933–949. https://doi.org/10.1093/brain/119.3.933

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Disentangling Syntax and Semantics in the Brain with Deep Networks. *Proceedings of the 38th International Conference on Machine Learning*. http://proceedings.mlr.press/v139/caucheteux21a/caucheteux21a.pdf

Chang, C., Leopold, D., Schölvinck, M., Mandelkow, H., Picchioni, D., Liu, X., Ye, F., Turchi, J., & Duyn, J. (2016). Tracking brain arousal fluctuations with fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(16), 4518–4523.

Chang, E. F., Kurteff, G., & Wilson, S. M. (2018). Selective Interference with Syntactic Encoding during Sentence Production by Direct Electrocortical Stimulation of the Inferior Frontal Gyrus. *Journal of Cognitive Neuroscience*, *30*(3), 411–420. https://doi.org/10.1162/jocn_a_01215

Chee, M. W. L., Tan, E. W. L., & Thiel, T. (1999). Mandarin and English Single Word Processing Studied with Functional Magnetic Resonance Imaging. *The Journal of Neuroscience*, *19*(8), 3050. https://doi.org/10.1523/JNEUROSCI.19-08-03050.1999

Chen, X., Affourtit, J., Ryskin, R., Regev, T. I., Norman-Haignere, S., Jouravlev, O., Malik-Moraleda, S., Kean, H., Varley, R., & Fedorenko, E. (2021). The human language system does not support music processing. *BioRxiv*, 2021.06.01.446439. https://doi.org/10.1101/2021.06.01.446439

Corina, D. P., Loudermilk, B. C., Detwiler, L., Martin, R. F., Brinkley, J. F., & Ojemann, G. (2010). Analysis of naming errors during cortical stimulation mapping: Implications for models of language representation. *Brain and Language*, *115*(2), 101–112. https://doi.org/10.1016/j.bandl.2010.04.001

Dapretto, M., & Bookheimer, S. Y. (1999). Form and Content: Dissociating Syntax and Semantics in Sentence Comprehension. *Neuron*, *24*(2), 427–432. https://doi.org/10.1016/S0896-6273(00)80855-7

Dasgupta, I., & Gershman, S. J. (2021). Memory as a Computational Resource. *Trends in Cognitive Sciences*, *25*(3), 240–251. https://doi.org/10.1016/j.tics.2020.12.008

de Bleser, R. (1987). From agrammatism to paragrammatism: German aphasiological traditions and grammatical disturbances. *Cognitive Neuropsychology*, *4*(2), 187–256. https://doi.org/10.1080/02643298708252039

Diachek, E., Blank, I., Siegelman, M., Affourtit, J., & Fedorenko, E. (2020). The Domain-General Multiple Demand (MD) Network Does Not Support Core Aspects of Language Comprehension: A Large-Scale fMRI Investigation. *Journal of Neuroscience*, *40*(23), 4536–4550. https://doi.org/10.1523/JNEUROSCI.2036-19.2020

Dick, F., Bates, E., Wulfeck, B., Utman, J. A., Dronkers, N., & Gernsbacher, M. A. (2001). Language deficits, localization, and grammar: Evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychological Review*, *108*(4), 759–788. https://doi.org/10.1037/0033-295X.108.4.759

Ding, J., Martin, R. C., Hamilton, A. C., & Schnur, T. T. (2020). Dissociation between frontal and temporal-parietal contributions to connected speech in acute stroke. *Brain*, *143*(3), 862–876. https://doi.org/10.1093/brain/awaa027

Duncan J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Duncan, J. (2013). The Structure of Cognition: Attentional Episodes in Mind and Brain. *Neuron*, *80*(1), 35–50. https://doi.org/10.1016/j.neuron.2013.09.015

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*(10), 475–483. https://doi.org/10.1016/S0166-2236(00)01633-7

Embick, D., Marantz, A., Miyashita, Y., O'Neil, W., & Sakai, K. L. (2000). A syntactic specialization for Broca's area. *Proceedings of the National Academy of Sciences*, *97*(11), 6150. https://doi.org/10.1073/pnas.100098897

Erdogan, S., Tong, Y., Hocke, L., Lindsey, K., & deB Frederick, B. (2016). Correcting for Blood Arrival Time in Global Mean Regression Enhances Functional Connectivity Analysis of Resting State fMRI-BOLD Signals. *Frontiers in Human Neuroscience*, *10*(311).

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual

Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108*(39), 16428–16433. https://doi.org/10.1073/pnas.1112937108

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *110*(41), 16616–16621. https://doi.org/10.1073/pnas.1315235110

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, *113*(41), E6256. https://doi.org/10.1073/pnas.1612132113

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, 104348. https://doi.org/10.1016/j.cognition.2020.104348

Fedorenko, E., & Blank, I. A. (2020). Broca's Area Is Not a Natural Kind. *Trends in Cognitive Sciences*, *24*(4), 270–284. https://doi.org/10.1016/j.tics.2020.01.001

Fedorenko, E., & Shain, C. (in press). Similarity of computations across domains does not imply shared implementation: The case of language comprehension. *Current Directions in Psychological Science.*

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, *18*(3), 120–126. https://doi.org/10.1016/j.tics.2013.12.006

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. https://doi.org/10.1111/1467-8721.00158

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T., & Crone, N. E. (2015). Redefining the role of Broca's area in speech. *Proceedings of the National Academy of Sciences*, *112*(9), 2871. https://doi.org/10.1073/pnas.1414491112

Fridriksson, J., Yourganov, G., Bonilha, L., Basilakos, A., Den Ouden, D.-B., & Rorden, C. (2016). Revealing the dual streams of speech processing. *Proceedings of the National Academy of Sciences*, *113*(52), 15108. https://doi.org/10.1073/pnas.1614038114

Friederici, A. D., Meyer, M., & Cramon, D. Y. von. (2000). Auditory Language Comprehension: An Event-Related fMRI Study on the Processing of Syntactic and Lexical Information. *Brain and Language*, *74*(2), 289–300. https://doi.org/10.1006/brln.2000.2313

Fuchs, S., Sock, R., & Laprie, Y. (2011). Speech is a very good example of a goal-directed organization of biological action. *Motor Control*, *15*(1), 2–4. https://doi.org/10.1123/mcj.15.1.2

Gibson, E., Sandberg, C., Fedorenko, E., Bergen, L., & Kiran, S. (2016). A rational inference approach to aphasic language comprehension. *Aphasiology*, *30*(11), 1341–1360. https://doi.org/10.1080/02687038.2015.1111994

Goodglass, H., & Geschwind, N. (1976). Language disturbance (aphasia). In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of Perception* (Vol. 7, pp. 389–428). Academic Press.

Goodglass, H., & Menn, L. (1985). Is Agrammatism a Unitary Phenomenon? In M.-L. Kean (Ed.), *Agrammatism* (pp. 1–26). Academic Press. https://doi.org/10.1016/B978-0-12-402830-2.50005-5

Guenther, F. H. (2016). *Neural Control of Speech*. MIT Press.

Hajnal, J., Myers, R., Oatridge, A., Schwieso, J., Young, I., & Bydder, G. (1994). Artifacts due to stimulus correlated motion in functional imaging of the brain. *Magnetic Resonance in Medicine*, *31*(3), 283–291.

Halai, A. D., Woollams, A. M., & Ralph, M. A. L. (2017). Using principal component analysis to capture individual differences within a unified neuropsychological model of chronic post-stroke aphasia: Revealing the unique neural correlates of speech fluency, phonology and semantics. *Cortex*, *86*, 275–289. https://doi.org/10.1016/j.cortex.2016.04.016

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, *19*(6), 304–313. https://doi.org/10.1016/j.tics.2015.04.006

He, H., Shin, D., & Liu, T. T. (2010). Resting state BOLD fluctuations in large draining veins are highly correlated with the global mean signal. *Proceedings of the 18th Annual Meeting of the ISMRM*, *3488*.

Hebart M.N., Dickter A.H., Kidder A., Kwok W.Y., Corriveau A., et al. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE* 14(10): e0223792. https://doi.org/10.1371/journal.pone.0223792

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1), 67–99. https://doi.org/10.1016/j.cognition.2003.10.011

Holmes, A. P., & Friston, K. J. (1998). Generalisability, Random Effects & Population Inference. *NeuroImage*, *7*(4, Part 2), S754. https://doi.org/10.1016/S1053-8119(18)31587-8

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. Zenodo. https://doi.org/10.5281/zenodo.1212303

Howard, D. (1985). Agrammatism. In S. Newman & R. Epstein (Eds.), *Perspectives in Dysphasiology* (pp. 1–31).

Hugdahl, K., Raichle, M. E., Mitra, A., & Specht, K. (2015). On the existence of a generalized non-specific task-dependent network. *Frontiers in Human Neuroscience*, *9*, 430. https://doi.org/10.3389/fnhum.2015.00430

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2007). Time course of semantic processes during sentence comprehension: An fMRI study. *NeuroImage*, *36*(3), 924–932. https://doi.org/10.1016/j.neuroimage.2007.03.059

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Indefrey, P. (2011). The Spatial and Temporal Signatures of Word Production Components: A Critical Update. *Frontiers in Psychology*, *2*, 255. https://doi.org/10.3389/fpsyg.2011.00255

Ivanova, A. A., Srikant, S., Sueoka, Y., Kean, H. H., Dhamala, R., O'Reilly, U.-M., Bers, M. U., & Fedorenko, E. (2020). Comprehension of computer code relies primarily on domain-general executive brain regions. *ELife*, *9*, e58906. https://doi.org/10.7554/eLife.58906

Jakobson, R. (1941/1968). *Child language, aphasia, and phonological universals*. Mouton.

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage*, *60*(4), 2357–2364. https://doi.org/10.1016/j.neuroimage.2012.02.055

Kaplan, E., Bekkering, H., Weintraub, S., & Goodglass, H. (1983). *The Boston naming test*. Lea & Febiger.

Keller, T. A., Carpenter, P. A., & Just, M. A. (2001). The Neural Bases of Sentence Comprehension: A fMRI Examination of Syntactic and Lexical Processing. *Cerebral Cortex*, *11*(3), 223–237. https://doi.org/10.1093/cercor/11.3.223

Kempen, G. (2000). Could grammatical encoding and grammatical decoding be subserved by the same processing module? *Behavioral and Brain Sciences*, *23*(1), 38–39. https://doi.org/10.1017/S0140525X00402396

Kempler, D., Curtiss, S., Metter, E. J., Jackson, C. A., & Hanson, W. R. (1991). Grammatical comprehension, aphasic syndromes and neuroimaging. *Journal of Neurolinguistics*, *6*(3), 301–318. https://doi.org/10.1016/0911-6044(91)90024-D

Kleiman, E. (2017). EMAtools: Data Management Tools for Real-Time Monitoring/Ecological Momentary Assessment Data. R package version 0.1.3. https://CRAN.R-project.org/package=EMAtools

Kojima, K., Brown, E. C., Matsuzaki, N., Rothermel, R., Fuerst, D., Shah, A., Mittal, S., Sood, S., & Asano, E. (2013). Gamma activity modulated by picture and auditory naming tasks: Intracranial recording in patients with focal epilepsy. *Clinical Neurophysiology*, *124*(9), 1737–1744. https://doi.org/10.1016/j.clinph.2013.01.030

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Kuperberg, G. R., Holcomb, P. J., Sitnikova, T., Greve, D., Dale, A. M., & Caplan, D. (2003). Distinct Patterns of Neural Modulation during the Processing of Conceptual and Syntactic Anomalies. *Journal of Cognitive Neuroscience*, *15*(2), 272–293. https://doi.org/10.1162/089892903321208204

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lee, D. K., Fedorenko, E., Simon, M. V., Curry, W. T., Nahed, B. V., Cahill, D. P., & Williams, Z. M. (2018). Neural encoding and production of functional morphemes in the posterior temporal lobe. *Nature Communications*, *9*(1), 1877. https://doi.org/10.1038/s41467-018-04235-3

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. MIT Press.

Levy, R. (2008). A Noisy-Channel Model of Human Sentence Comprehension under Uncertain Input. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243. https://aclanthology.org/D08-1025

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye Movement Evidence that Readers Maintain and Act on Uncertainty about Past Linguistic Input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090. https://www.pnas.org/content/106/50/21086

Lipkin, B., Small, H., Affourtit, J., Mineroff, Z., Nieto-Castañón, A., & Fedorenko, E. (in prep). In defense of individual-level functional neural markers: Evidence from large-scale fMRI datasets of functional 'localizers' for the language and the Multiple Demand networks.

Long, M. A., Katlowitz, K. A., Svirsky, M. A., Clary, R. C., Byun, T. M., Majaj, N., Oya, H., Howard, M. A., & Greenlee, J. D. W. (2016). Functional Segregation of Cortical Regions Underlying Speech Timing and Articulation. *Neuron*, *89*(6), 1187–1193. https://doi.org/10.1016/j.neuron.2016.01.032

Longcamp, M., Lagarrigue, A., Nazarian, B., Roth, M., Anton, J.-L., Alario, F.-X., & Velay, J.-L. (2014). Functional specificity in the motor system: Evidence from coupled fMRI and kinematic recordings during letter and digit writing. *Human Brain Mapping*, *35*(12), 6077–6087. https://doi.org/10.1002/hbm.22606

Lu, J., Zhao, Z., Zhang, J., Wu, B., Zhu, Y., Chang, E. F., Wu, J., Duffau, H., & Berger, M. S. (2021). Functional maps of direct electrical stimulation-induced speech arrest and anomia: A multicentre retrospective study. *Brain*, *144*(8), 2541–2553. https://doi.org/10.1093/brain/awab125

MacSweeney, M., Capek, C. M., Campbell, R., & Woll, B. (2008). The signing brain: The neurobiology of sign language. *Trends in Cognitive Sciences*, *12*(11), 432–440. https://doi.org/10.1016/j.tics.2008.07.010

Mahowald, K., & Fedorenko, E. (2016). Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *NeuroImage*, *139*, 74–93. https://doi.org/10.1016/j.neuroimage.2016.05.073

Marslen-Wilson, W. (1973). Linguistic Structure and Speech Shadowing at Very Short Latencies. *Nature*, *244*(5417), 522–523. https://doi.org/10.1038/244522a0

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared Language: Overlap and Segregation of the Neuronal Infrastructure for Speaking and Listening Revealed by Functional MRI. *Psychological Science*, *22*(9), 1173–1182. https://doi.org/10.1177/0956797611418347

Menn, L., & Matthei, E. (2011). The two-lexicon approach of child phonology: Looking back, looking ahead. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 211–248). York Press.

Mesulam, M. M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology*, *28*(5), 597–613. https://doi.org/10.1002/ana.410280502

Mesulam, M.-M., Thompson, C. K., Weintraub, S., & Rogalski, E. J. (2015). The Wernicke conundrum and the anatomy of language comprehension in primary progressive aphasia. *Brain*, *138*(8), 2423–2437. https://doi.org/10.1093/brain/awv154

Miceli, G., Mazzucchi, A., Menn, L., & Goodglass, H. (1983). Contrasting cases of Italian agrammatic aphasia without comprehension disorder. *Brain and Language*, *19*(1), 65–97. https://doi.org/10.1016/0093-934X(83)90056-1

Mineroff, Z., Blank, I. A., Mahowald, K., & Fedorenko, E. (2018). A robust dissociation among the language, multiple demand, and default mode networks: Evidence from inter-region correlations in effect size. *Neuropsychologia*, *119*, 501–511. https://doi.org/10.1016/j.neuropsychologia.2018.09.011

Miyake, A., Carpenter, P. A., & Just, M. A. (1994). A capacity approach to syntactic comprehension disorders: Making normal adults perform like aphasic patients. *Cognitive Neuropsychology*, *11*(6), 671–717. https://doi.org/10.1080/02643299408251989

Nespoulous, J.-L., Dordain, M., Perron, C., Ska, B., Bub, D., Caplan, D., Mehler, J., & Lecours, A. R. (1988). Agrammatism in sentence production without comprehension deficits: Reduced availability of syntactic structures and/or of grammatical morphemes? A case study. *Brain and Language*, *33*(2), 273–295. https://doi.org/10.1016/0093-934X(88)90069-7

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

Ni, W., Constable, R. T., Mencl, W. E., Pugh, K. R., Fulbright, R. K., Shaywitz, S. E., Shaywitz, B. A., Gore, J. C., & Shankweiler, D. (2000). An Event-related Neuroimaging Study Distinguishing Form and Content in Sentence Processing. *Journal of Cognitive Neuroscience*, *12*(1), 120–133. https://doi.org/10.1162/08989290051137648

Nieto-Castañón, A., & Fedorenko, E. (2012). Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage*, *63*(3), 1646–1669. https://doi.org/10.1016/j.neuroimage.2012.06.065

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

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

Parisi, D. (1987). Grammatical disturbances of speech production. In M. Coltheart, G. Sartori, & R. Job (Eds.), *The Cognitive Neuropsychology of Language*. Erlbaum.

Paunov, A. M., Blank, I. A., & Fedorenko, E. (2019). Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. *Journal of Neurophysiology*, *121*(4), 1244–1265. https://doi.org/10.1152/jn.00619.2018

Plummer, B.A., Wang, L., Cervantes, C.M., et al. (2017). Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *Int J Comput Vis* 123**,** 74–93. https://doi.org/10.1007/s11263-016-0965-7

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron*, *72*(5), 692–697. Cell Press. https://doi.org/10.1016/j.neuron.2011.11.001

Potter, M. (2012). Conceptual Short Term Memory in Perception and Thought. *Frontiers in Psychology, 3*, 113. https://doi.org/10.3389/fpsyg.2012.00113

Power, J., Schlaggar, B., & Petersen, S. (2015). Recent progress and outstanding issues in motion correction in resting state fMRI. *NeuroImage*, *107*, 536–551.

Python, G., Glize, B., & Laganaro, M. (2018). The involvement of left inferior frontal and middle temporal cortices in word production unveiled by greater facilitation effects following brain damage. *Neuropsychologia*, *121*, 122–134. https://doi.org/10.1016/j.neuropsychologia.2018.10.026

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

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology*, 55A, 1339-1362.

Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and Invariant Neural Responses to Spoken and Written Narratives. *Journal of Neuroscience*, *33*(40), 15978–15988. https://doi.org/10.1523/JNEUROSCI.1580-13.2013

Regev, T. I., Affourtit, J., Chen, X., Schipper, A. E., Bergen, L., Mahowald, K., & Fedorenko, E. (2021). High-level language brain regions are sensitive to sub-lexical regularities. *BioRxiv*, 2021.06.11.447786. https://doi.org/10.1101/2021.06.11.447786

Rodd, J. M., Vitello, S., Woollams, A. M., & Adank, P. (2015). Localising semantic and syntactic processing in spoken and written language comprehension: An Activation Likelihood Estimation meta-analysis. *Brain and Language*, *141*, 89–102. https://doi.org/10.1016/j.bandl.2014.11.012

Röder, B., Stock, O., Neville, H., Bien, S., & Rösler, F. (2002). Brain Activation Modulated by the Comprehension of Normal and Pseudo-word Sentences of Different Processing

Demands: A Functional Magnetic Resonance Imaging Study. *NeuroImage*, *15*(4), 1003–1014. https://doi.org/10.1006/nimg.2001.1026

Roux, F.-E., Dufor, O., Giussani, C., Wamain, Y., Draper, L., Longcamp, M., & Démonet, J.-F. (2009). The graphemic/motor frontal area Exner's area revisited. *Annals of Neurology*, *66*(4), 537–545. https://doi.org/10.1002/ana.21804

Schnur, T. T., Schwartz, M. F., Kimberg, D. Y., Hirshorn, E., Coslett, H. B., & Thompson-Schill, S. L. (2009). Localizing interference during naming: Convergent neuroimaging and neuropsychological evidence for the function of Broca's area. *Proceedings of the National Academy of Sciences*, *106*(1), 322. https://doi.org/10.1073/pnas.0805874106

Schölvinck, M., Maier, A., Ye, F., Duyn, J., & Leopold, D. (2010). Neural basis of global resting-state fMRI activity. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(22), 10238–10243.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., & Fedorenko, E. (2020). Artificial Neural Networks Accurately Predict Language Processing in the Brain. *BioRxiv*, 2020.06.26.174482. https://doi.org/10.1101/2020.06.26.174482

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

Shain, C., Blank, I. A., Schijndel, M. van, Schuler, W., & Fedorenko, E. (2020). FMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Shain, C., Kean, H., Affourtit, J., Siegelman, M., Mollica, F., & Fedorenko, E. (in prep). Evidence against syntax abstractness and autonomy in the language network: A failure to replicate critical aspects of Pallier, Devauchelle, and Dehaene (2011, PNAS).

Shashidhara, S., Mitchell, D. J., Erez, Y., & Duncan, J. (2019). Progressive Recruitment of the Frontoparietal Multiple-demand System with Increased Task Complexity, Time Pressure, and Reward. *Journal of Cognitive Neuroscience*, *31*(11), 1617–1630. https://doi.org/10.1162/jocn_a_01440

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, *111*(43), E4687. https://doi.org/10.1073/pnas.1323812111

Stark, J. A., & Dressler, W. U. (1990). Agrammatism in German: Two case studies. In L. Menn & L. K. Obler (Eds.), *Agrammatic Aphasia: A cross-language narrative sourcebook*. John Benjamins Publishing Company.

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 14954–14964. https://arxiv.org/abs/1905.11833

Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C., & Dehaene, S. (2012). A Temporal Bottleneck in the Language Comprehension Network. *Journal of Neuroscience*, *32*(26), 9089–9102. https://doi.org/10.1523/JNEUROSCI.5685-11.2012

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., Smith, N., Gibson, E., & Fedorenko, E. (2021). Incremental Language Comprehension Difficulty

Predicts Activity in the Language Network but Not the Multiple Demand Network. *Cerebral Cortex*, *31*(9), 4006–4023. https://doi.org/10.1093/cercor/bhab065

Willems, R. M., der Haegen, L. V., Fisher, S. E., & Francks, C. (2014). On the other hand: Including left-handers in cognitive neuroscience and neurogenetics. *Nature Reviews Neuroscience*, *15*(3), 193–201. https://doi.org/10.1038/nrn3679

Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., & Shenoy, K. V. (2021). High-performance brain-to-text communication via handwriting. *Nature*, *593*(7858), 249–254. https://doi.org/10.1038/s41586-021-03506-2

Wilson, S. M., & Saygın, A. P. (2004). Grammaticality Judgment in Aphasia: Deficits Are Not Specific to Syntactic Structures, Aphasic Syndromes, or Lesion Sites. *Journal of Cognitive Neuroscience*, *16*(2), 238–252. https://doi.org/10.1162/089892904322984535

Wong, C., Olafsson, V., Tal, O., & Liu, T. (2013). The amplitude of the resting-state fMRI global signal is related to EEG vigilance measures. *NeuroImage*, *83*, 989–990.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association of Computational Linguistics* 2, 67-78. http://dx.doi.org/10.1162/tacl_a_00166