Distributed Sensitivity to Syntax and Semantics throughout the Language Network

Cory Shain[1*], Hope Kean[1*], Colton Casto[1], Benjamin Lipkin[1], Josef Affourtit[1], Matthew Siegelman[1,2], Francis Mollica[3†], and Evelina Fedorenko[1,4†]

Abstract

■ Human language is expressive because it is *compositional*: The meaning of a sentence (semantics) can be inferred from its structure (syntax). It is commonly believed that language syntax and semantics are processed by distinct brain regions. Here, we revisit this claim using precision fMRI methods to capture separation or overlap of function in the brains of individual participants. Contrary to prior claims, we find distributed sensitivity to both syntax and semantics throughout a broad frontotemporal brain network. Our results join a growing body of evidence for an integrated network for language in the human brain within which internal specialization is primarily a matter of degree rather than kind, in contrast with influential proposals that advocate distinct specialization of different brain areas for different types of linguistic functions. ■

INTRODUCTION

Human language is a powerful medium for communicating complex thoughts. This power comes from the compositional structure of language (Chomsky, 1965): Meaning is encoded not only by individual words but also by the form and sequential arrangement of those words. For example, the sentence *There are octopuses inside the bathtub!* is (probably) unfamiliar to the reader and also (probably) expresses a meaning with which the reader has no direct experience. Yet novel meanings are recoverable from novel sentences thanks to the systematic relationship between a sentence's form and its meaning. This principle even extends to unfamiliar words: When we read *There are blickets inside the dax!*, we can infer that the blickets and the dax are in a containment relationship and have certain other properties (e.g., a *blicket* is countable and a *dax* can contain something), even if we do not know the meanings of the words themselves. Thus, the expressive power of language derives from its factorization into meaning (*semantics*) versus form (the sentence's grammatical structure or *syntax*).

Many models of the neurobiology of language posit a similar factorization at the level of large-scale brain areas (e.g., posterior temporal vs. inferior frontal areas)—such that some areas are "syntactic hubs" that selectively represent and process the grammatical structure of sentences, whereas others are "semantic hubs" that selectively represent and process the meanings of words and/or phrases/sentences—albeit with disagreement as to the precise locations of these functions in the brain (Friederici, 2017; Duffau, Moritz-Gasser, & Mandonnet, 2014; Bornkessel-Schlesewsky & Schlesewsky, 2013; Hickok & Poeppel, 2007; Hagoort, 2005; Frazier, 1987). If true, this division of linguistic labor would have fundamental implications for the organization and evolutionary origins of human cognition: Brain circuits that implement the abstract combinatorics needed for syntactic processing could be recruited in service of other cognitive functions that have similar hierarchical structure to language (e.g., mathematics, music, and action planning; Koechlin & Jubault, 2006; Patel, 2003; Lashley, 1951), and they may find their origins in changes to brain anatomy that enabled algebraic thought, including language (Dehaene, Al Roumi, Lakretz, Planton, & Sablé-Meyer, 2022; Dehaene, Meyniel, Wacongne, Wang, & Pallier, 2015). One important source of evidence in favor of the spatial separability of syntax and semantics has been a landmark study by Pallier, Devauchelle, and Dehaene (Pallier, Devauchelle, & Dehaene, 2011, henceforth PDD), who argued based on fMRI evidence for a dissociation between brain areas that selectively represent and process syntax and areas that selectively represent and process *lexical* (word-level) and *combinatorial* (phrase-level) semantics. PDD's claims have informed theories of cognition, brain function, and evolution, especially those that posit neural circuits dedicated to abstract combinatorics (e.g., Dehaene et al., 2015, 2022; Bornkessel-Schlesewsky, Schlesewsky, Small, & Rauschecker, 2015; Bolhuis, Tattersall, Chomsky, & Berwick, 2014; Fitch, 2014; Petkov & Jarvis, 2012).

[1]Massachusetts Institute of Technology, [2]Columbia University, [3]University of Edinburgh, [4]Harvard University
*Equal contribution.
†Co-senior authorship.

Journal of Cognitive Neuroscience 36:7, pp. 1427–1471
https://doi.org/10.1162/jocn_a_02164

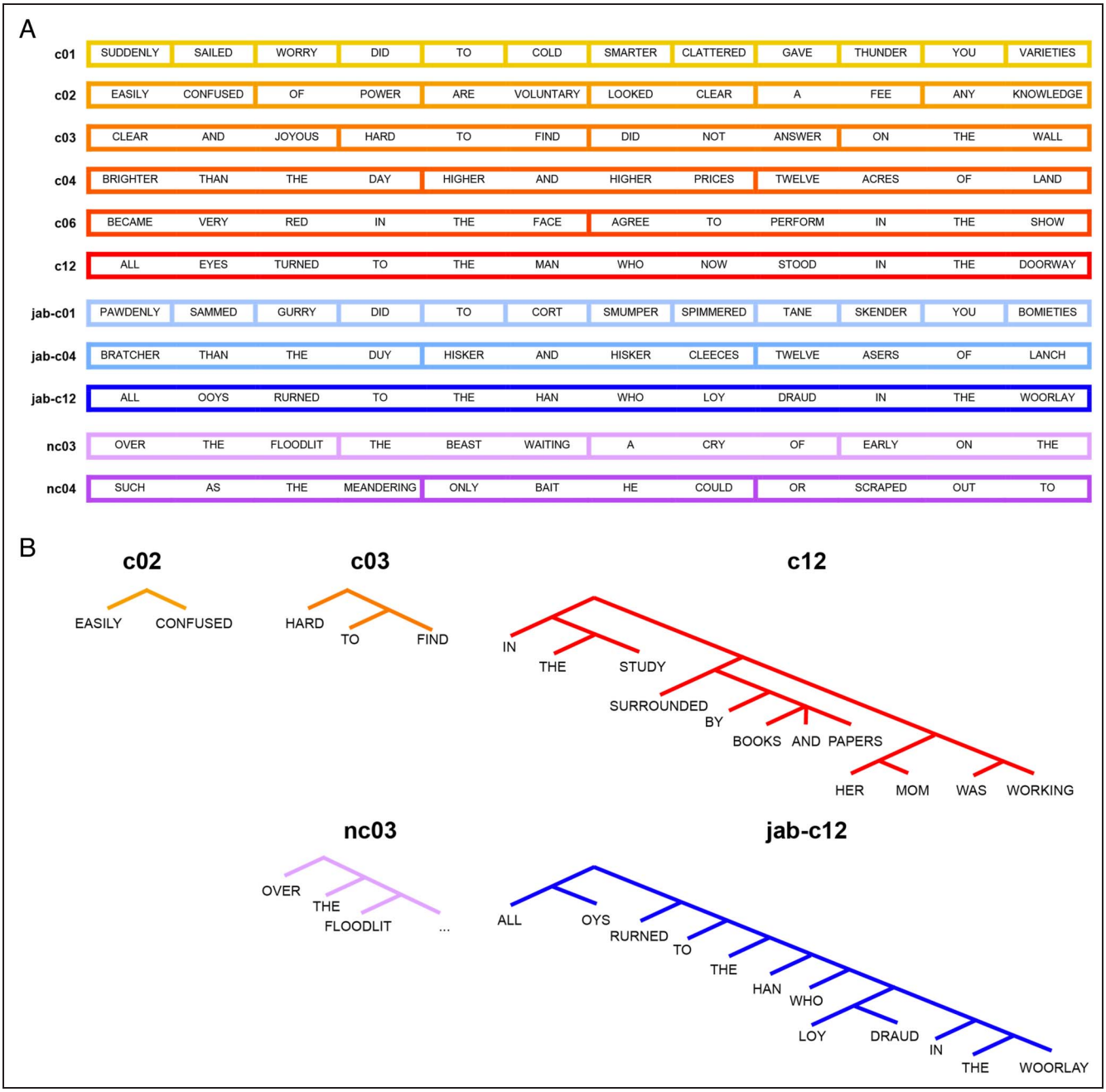

Figure 1. (A) Examples of stimuli (Experiments 1 and 2) across conditions (from one-word chunks, c01, to 12-word chunks, c12), with real-word constituent conditions shown in warm colors, Jabberwocky constituent conditions shown in blues, and real-word, nonconstituent conditions shown in purples. Note the morphosyntactic parallelism between the pairs c01/jab-c01, c04/jab-c04, and c12/jab-c12. (B) Visualization of constituent structure of representative chunks. In a phrase-structure grammar, a *constituent* is the entire sequence of words that is dominated by a branching node in the tree. In the c02 condition, there is exactly one constituent ("easily confused"), whereas in the c12 condition, many constituents are nested (e.g., "in the study" is a constituent nested within the entire sentence, which is itself a constituent). The same kind of nested constituency structure is implicit in the Jabberwocky condition (jab-c12), although most of the words are meaningless. By contrast, in the nonconstituent condition (nc03), the three words ("over the floodlit") do not form a constituent, because the only node in the tree that dominates all of them (the top-most node) implicitly contains at least one additional missing word (the noun modified by "floodlit").

In PDD's paradigm (Figure 1A), participants read 12-word stimuli presented one word at a time. These stimuli were internally composed of "chunks" (our terminology) of locally coherent connected words. The chunks varied parametrically in length. At one extreme, a stimulus contained 12 concatenated (one-word) chunks (condition "c01" in Figure 1A), and at the other, a stimulus contained a single 12-word chunk (condition "c12" in Figure 1A). In the intermediate conditions, the stimuli contained concatenated chunks of different lengths: six 2-word

chunks (c02), four 3-word chunks (c03), three 4-word chunks (c04), or two 6-word chunks (c06). The chunks in these conditions always formed valid syntactic *constituents*, that is, a complete phrase in a hierarchical representation of the sentence's grammatical structure (i.e., a *parse tree*; see Figure 1B). PDD hypothesized that language processing requires the comprehender to maintain an increasingly complex representation of structure (i.e., a unified syntactic and/or semantic representation of the word sequence) as each new word is processed, and that this increased representational complexity will correspond to an increase in overall neuronal activity in conditions with longer constituents (given that they express a more complex phrasal structure; see Figure 1B). To investigate the abstractness of syntactic representations, a "Jabberwocky" version of each condition (e.g., jab-c01, jab-c12) was created by replacing the content words (nouns, verbs, adjectives, and adverbs) with word-like nonwords (pseudowords), but preserving the syntactic "frame," that is, function words like articles and auxiliaries, and functional morphological endings (e.g., *higher and higher prices* > *hisker and hisker cleeces*).

This design targets three potentially dissociable dimensions of linguistic representation, each of which could be either present or absent in a given brain region's response. *First*, **lexical semantics** (learned conventions about individual word meanings) is fully present only in the real-word conditions. Although Jabberwocky conditions may allow some aspects of lexical meaning to be inferred (rather than recalled)—for example, via pseudowords' form (Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016) or context (Li, 1988)—the content of any such inferences should be consistent with a broader range of meanings on average than learned word meanings; otherwise, vocabulary learning would be unnecessary for language comprehension. Second, **syntax** (the grammatical structure of the sentence as reflected in the forms and sequential ordering of words) is present in proportion to chunk length. When chunk length is 1 (i.e., a word list), syntactic demands should be limited to the processing of word-internal features (e.g., suffixes marking tense or plurality), with no possibility of integrating words into larger structures (e.g., parse trees). As chunk length increases, the constituents formed by each chunk increase in complexity, which may impose additional processing demands related to syntactic tree construction (e.g., hypothesizing new nodes in the tree and new grammatical dependencies to preceding words in the chunk). Importantly, because the surface cues that are needed to construct abstract trees (prefixes, suffixes, and function words) are present in both the real-word and Jabberwocky conditions, the demands associated with tree construction should be similarly modulated by chunk length in both types of conditions. PDD also identified specific effects of phrasal constituency by including *nonconstituents* (nc) versions of the c03 and c04 conditions, that is, three- and four-word coherent sequences of real words that do not form complete syntactic phrases (e.g., *over the floodlit*; Figure 1B). Third, similar to lexical semantics, **combinatorial semantics** (the composite meaning denoted by the chunk) is fully present only in the real-word conditions. Although certain abstract properties of meaning are directly encoded by syntax (e.g., plurality in *cleeces* is both a syntactic and a semantic property) and thus present in the Jabberwocky conditions, these meanings are impoverished relative to real-word conditions in ways that go beyond the mere absence of lexical semantics. The combinatorial semantics of real-word items enable the construction of detailed mental models of meaning that can themselves inform other inferences not directly stated in language. Returning to the example above, the sentence *There are octopuses inside the bathtub!* may license the inference that there is also water in the bathtub with greater confidence than the sentence *There are blickets inside the dax!* may license the inference that there is also water in the dax, because the real-word sentence permits more specific connections to conceptual knowledge (e.g., that octopuses are aquatic animals, that bathtubs hold water). Furthermore, similar to syntax, combinatorial semantics is present in proportion to chunk length. When chunk length is 1, the meanings of nearby words cannot be unified into a larger whole. As chunk length increases, the mental representations of entities can be refined (e.g., via adjectives and prepositional phrases) and relations between entities can be hypothesized (e.g., by recognizing the subjects and objects of verbs).

These three linguistic dimensions (lexical semantics, syntax, and combinatorial semantics) coordinate to produce the compositional power of human language discussed above. Words with conventionalized meanings (lexical semantics) can be *composed* (i.e., unified into a single syntactic and/or semantic representation) by a shared system of rules (syntax) so as to lead systematically to shared representations of more complex meanings (combinatorial semantics) that may themselves be novel, or lack a single conventionalized expression. Differences in brain response to word sequences as a function of whether those sequences can be composed can thus shed light on the nature of composition during language comprehension (e.g., Vandenberghe, Nobre, & Price, 2002), including whether different components of composition are implemented by different brain areas. For convenience, our assumed definitions of key terms in the foregoing discussion are provided in Table 1.

PDD's design therefore gives rise to the eight hypothetical response profiles depicted in Figure 2. For example, a selectively syntactic region (−Lex, +Syn, –CombSem) should respond identically across real-word and Jabberwocky conditions; a selectively combinatorial-semantic region (−Lex, –Syn, +CombSem) should show a length effect (stronger responses to longer chunks) only in the real-word conditions; and a combined lexical-semantic, syntactic, and combinatorial-semantic region (+Lex, +Syn, +CombSem) should show length effects in both real-word and Jabberwocky conditions, with a stronger length effect in the real-word conditions. PDD's design

Table 1. Definitions of Key Terms as Used in This Article

Syntax	Grammatical properties that govern the form and arrangement of words in sentences. Within *syntax*, we include phrase structure (e.g., a determiner such as *the* followed by a noun such as *cat* can form a noun phrase *the cat*), grammatical relations (e.g., the phrase *the cat* can be the subject of a verb, e.g., *sleeps*), and affixation patterns that reflect these relations (e.g., the suffix *-s* of *sleeps* indicates that the subject is singular).
Lexical semantics	Meanings of words. Within *lexical semantics*, we include all learned information about the concepts, properties, affordances, and usage patterns (e.g., social register) associated with words in a language. We exclude any aspect of meaning that can be inferred from the form of the word alone (e.g., plurality, which is often grammatically marked in English).
Combinatorial semantics	Meanings of multiword phrases. Within *combinatorial semantics*, we include any aspect of meaning that is not directly conveyed by the words considered in isolation. This includes meanings inferred from the syntactic arrangement of words (e.g., merging the representations of the meanings of *the cat* and *sleeps* to yield a representation of the proposition *the cat sleeps* in some system of formal logic, e.g., Church, 1932) and the meanings of conventionalized collocations (e.g., *let the cat out of the bag* to mean "accidentally reveal a secret").
Structure	Any property of multiword phrases that is not directly conveyed by the words considered in isolation, including both grammar and meaning. We thus use *structure* as a cover term for both syntax and combinatorial semantics. In our experiments, structure (more precisely, structural complexity) is modulated by chunk length.
Composition	Any mental process that infers (syntactic and/or conceptual) structure from sequences of words.
Constituent	The full word sequence dominated by a single node in a hierarchical representation of a sentence's phrase structure (see Figure 1B).
Chunk	A contiguous sequence of (pseudo)words that can be syntactically and/or semantically composed.
Length effect	An increase in a brain region's BOLD response as a function of chunk length.

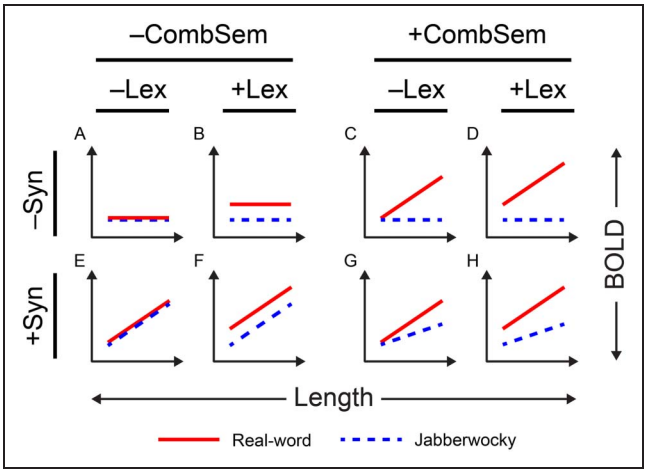

Figure 2. Hypothetical outcomes of PDD's experiment under different sensitivities to lexical-semantic (±Lex), syntactic (±Syn), and combinatorial-semantic (±CombSem) dimensions of language. Lexical-semantic processing (+Lex) predicts a larger overall response to real-word conditions than to Jabberwocky conditions, shifting all estimates for real-word conditions upward. Syntactic processing (+Syn) predicts an increase in response to chunk length (*x* axis) in both the real-word and Jabberwocky conditions. Combinatorial-semantic processing (+CombSem) predicts a greater response to chunk length in the real-word conditions than the Jabberwocky conditions. These predictions combine to yield eight logically possible response profiles, many of which can be distinguished by testing for differences by condition type between the intercept (overall response) and/or slope (strength of response to chunk length).

therefore permits empirical discrimination of different logically possible patterns of (in)sensitivity to lexical-semantic, syntactic, and combinatorial-semantic dimensions of language, with major implications for our understanding of the neural substrates that enable language comprehension.

PDD reported three key findings of relevance to the neural substrates of syntactic and semantic processing. Finding 1: Inferior frontal and posterior temporal language regions in the left hemisphere (LH) responded more strongly to longer constituents, even in the Jabberwocky conditions. Finding 2: The *slope* of this increase with chunk length was indistinguishable in these areas between the real-word and Jabberwocky conditions. The impact of this finding was plausibly enhanced by the additional apparent absence of a difference in *intercept* between conditions, such that the overall response profiles in these regions were nearly identical in the two types of conditions (similar to the selectively syntactic, −Lex, +Syn, –CombSem, profile in Figure 2). Finding 3: By contrast, in anterior temporal and temporoparietal language regions, activation increased with chunk length in the real-word conditions but not the Jabberwocky conditions, with a significant difference in slope between the two condition types (similar to the –Syn, +CombSem profiles in Figure 2). These findings have been reinforced by other studies showing syntactic/semantic dissociations with a similar topography to that reported by

PDD (e.g., Matchin, Hammerly, & Lau, 2017; Goucha & Friederici, 2015).

In addition to their support for neurobiological effects of syntax in general (Finding 1), these findings have had a major influence on thinking about the division of labor within the human language system, which we group into two broad claims that were made directly by PDD or attributed to them by subsequent work.

- Syntactic Hubs: Finding 2 has been taken to support the existence of abstract-syntactic hubs in inferior frontal and posterior temporal cortex (Dehaene, 2019; Nelson et al., 2017; Hertrich, Dietrich, & Ackermann, 2016; Pattamadilok, Dehaene, & Pallier, 2016; Dehaene et al., 2015; Wang, Uhrig, Jarraya, & Dehaene, 2015; Pallier et al., 2011). Because PDD reported qualitatively identical response profiles in these regions for real-word and Jabberwocky conditions, prior invocations of this empirical finding are often ambiguous between a strong form in which these hubs exclusively encode abstract combinatorics—with no reference to lexical- or combinatorial-semantic content (Hertrich et al., 2016; Dehaene et al., 2015; Kempen, 2014; profile –Lex, +Syn, –CombSem in Figure 2)—and a weaker form in which these hubs do not encode combinatorial semantics, but may nonetheless respond more strongly to real words than pseudowords overall (Matchin et al., 2017; profile +Lex, +Syn, –CombSem in Figure 2).
- Lexico-Semantic Hubs: Finding 3 has been taken to support a selective role for anterior temporal and temporoparietal areas in lexical- and combinatorial-semantic processing (Frankland & Greene, 2020; Zaccarella, Schell, & Friederici, 2017; Bautista & Wilson, 2016; Friston & Buzsáki, 2016; Skeide, Brauer, & Friederici, 2016; Bornkessel-Schlesewsky et al., 2015; Zaccarella & Friederici, 2015; Wilson et al., 2014; Bornkessel-Schlesewsky & Schlesewsky, 2013; Pallier et al., 2011; profile +Lex, –Syn, +CombSem in Figure 2). PDD use the term *lexico-semantic* to characterize these areas, and we follow this terminology when discussing PDD's (and related) claims. For elaboration on the ways in which PDD's study has influenced subsequent thinking about the neurobiology of language, see Appendix 1.

However, these claims now face empirical and methodological objections. Empirically, the existence of syntactic hubs (or, at least, the strong form of this claim) has been challenged by evidence of lexical processing in the inferior frontal and posterior temporal areas identified by PDD as abstract syntactic hubs (e.g., Matchin et al., 2017; Fedorenko, Duncan, & Kanwisher, 2012; Fedorenko, Nieto-Castañón, & Kanwisher, 2012; Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010; Rodd, Davis, & Johnsrude, 2005), and the existence of lexico-semantic hubs has been challenged by evidence of sensitivity to structure in Jabberwocky materials in anterior temporal regions argued by PDD to be insensitive to such effects (e.g., Fedorenko, Duncan, et al., 2012; Fedorenko, Nieto-Castañón, et al., 2012; Fedorenko et al., 2010; Rogalsky & Hickok, 2009; Humphries, Binder, Medler, & Liebenthal, 2006). These prior studies raise concerns about the robustness and replicability of PDD's reported pattern. Methodologically, some of the choices in PDD's design and analyses are problematic. First, PDD used a between-subjects design to compare the real-word and Jabberwocky conditions (thus simultaneously varying both the sample of participants and the condition), although this manipulation is feasible to perform in a within-subject design that avoids this confound. Because individuals and, by extension, groups of individuals vary along numerous trait and state dimensions that are known to affect neural responses (e.g., Chen, Saad, Britton, Pine, & Cox, 2013; Hariri, 2009; Holmes & Friston, 1998), the magnitudes of neural responses in two groups cannot be confidently attributed to differences/similarities between conditions. Second, PDD used the same data both to define the ROIs and to quantify their responses, introducing circularity (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). Third, PDD relied on traditional fMRI group analyses (Holmes & Friston, 1998), which assume voxel-wise correspondence across individual brains. Ample evidence now exists for substantial interindividual variability in the precise locations of functional areas in the association cortex (e.g., Vázquez-Rodríguez et al., 2019; Frost & Goebel, 2012; Tahmasebi et al., 2012), including in the language network (e.g., Mahowald & Fedorenko, 2016; Fedorenko et al., 2010). Given that some of PDD's claims rely on not finding certain effects in certain brain regions, the choice of traditional group analyses, which suffer from low sensitivity (Nieto-Castañón & Fedorenko, 2012), is suboptimal. We stress that PDD's approach and claims were reasonable for the time and that some of the concerns above arise from empirical findings or methodological insights that were contemporaneous or subsequent to PDD's publication date. However, because PDD's findings continue to exert substantial influence, it is important to consider them in light of subsequent developments.

Motivated by these concerns, and in line with current emphasis in the field on robustness and replicability (Gilmore, Diaz, Wyble, & Yarkoni, 2017; Yarkoni & Westfall, 2017; Open Science Collaboration, 2015; Makel & Plucker, 2014; Simons, 2014; Pashler & Wagenmakers, 2012), we conduct three fMRI experiments (across $n = 75$ participants) that constitute the closest effort to date to replicate PDD's original study while addressing the methodological issues above. First, we use a strictly within-subject design. Second, we use independent data to define the ROIs and to quantify their responses to the critical conditions. And third, we define ROIs functionally in individual brains (e.g., Fedorenko, 2021; Fedorenko et al., 2010; Saxe, Brett, & Kanwisher, 2006), which has been shown to yield higher sensitivity

and higher functional resolution (e.g., Braga, DiNicola, Becker, & Buckner, 2020; Shashidhara, Spronkers, & Erez, 2020; Fedorenko, Duncan, et al., 2012; Nieto-Castañón & Fedorenko, 2012).

To foreshadow our results, we strongly replicate PDD's key discovery of a basic chunk length effect in all experiments (see Giglio, Ostarek, Weber, & Hagoort, 2022, for another recent replication by another research group): Activity in multiple language areas increases parametrically with the increasing length of linguistic context, even in the absence of lexical content. However, our results challenge the existence of both syntactic and semantic hubs. In particular, (a) all regions of the language network (except the TPJ / angular gyrus language region) show a length effect in Jabberwocky conditions; (b) all language regions show an effect of "lexicality," with real-word conditions eliciting stronger responses than Jabberwocky conditions; and (c) all language regions but the PostTemp language region show a length by lexicality interaction whereby the length effect is stronger in the real-word conditions compared with Jabberwocky conditions. We further show that these length effects do not critically depend on syntactic constituency per se but rather on the length of contiguous coherent text, which undermines PDD's claim that syntactic constituency critically drives the length effect.

These findings challenge a bifurcation of the language system into discrete syntactic and lexico-semantic components. Our results instead join a growing body of evidence for an integrated network for language in the human brain (Malik-Moraleda et al., 2022; Braga et al., 2020; Scott, Gallée, & Fedorenko, 2017; Fedorenko et al., 2010) within which internal specialization is primarily a matter of degree rather than kind (Blank & Fedorenko, 2020; Fedorenko & Blank, 2020; Fedorenko, Behr, & Kanwisher, 2011; Fedorenko, Blank, Siegelman, & Mineroff, 2020; Keller, Carpenter, & Just, 2001, see Fedorenko, Ivanova, & Regev, 2024, for review) in contrast with influential proposals that posit a sharp separation between different types of linguistic representations and processes (Friederici, 2017; Bornkessel-Schlesewsky & Schlesewsky, 2013; Hagoort, 2005; Hickok & Poeppel, 2007).

METHODS

This study consists of three experiments. Experiment 1 focuses on the real-word conditions from PDD and attempts to replicate the basic length effect in the language network's response. Experiment 2 additionally includes Jabberwocky conditions to test PDD's critical theoretical claim: that a subset of the language network implements abstract, content-independent, syntactic processing. Experiment 3 targets the centrality of syntactic constituency by investigating length effects using chunks that overwhelmingly do not form syntactic constituents.

Participants

Seventy-four unique individuals (age 18–38 years, 39 female participants) participated for payment (Experiment 1: $n = 15$; Experiment 2: $n = 40$, Experiment 3: $n = 20$; one individual participated in both Experiment 2 and Experiment 3, on separate days). The same number of participants (40) were included in our key replication of PDD (Experiment 2) as in PDD's original study. All but three participants were right-handed—as determined by the Edinburgh Handedness Inventory (Oldfield, 1971), or self-report. All participants were native (age of first exposure < 10 years old) or highly proficient ($n = 3$) speakers of English (see Malik-Moraleda et al., 2022, for evidence that the language system of highly proficient speakers is similar to that of native speakers). All participants gave informed written consent in accordance with the requirements of Massachusetts Institute of Technology's (MIT) Committee on the Use of Humans as Experimental Subjects. Each participant completed a language localizer task (Fedorenko et al., 2010) and a critical task.

Critical Task: Design

The design of Experiments 1 and 2 followed PDD but used English materials available at https://osf.io/fduve/ (the original experiments were carried out in French). In particular, participants were presented with same-length stimuli (sequences of 12 words/nonwords), and the internal composition of these stimuli varied across conditions. The conditions in Experiment 1 were similar to PDD's real-word conditions, except they did not include the three-word constituent condition. Experiment 2 included three types of experimental manipulation that directly follow PDD's original design: (a) six real-word conditions: a sequence of 12 unconnected words (i.e., constituents of length 1: c01; here and elsewhere, our condition name abbreviations are similar to those in PDD), six 2-word constituents (c02), four 3-word constituents (c03), three 4-word constituents (c04), two 6-word constituents (c06), and a 12-word sentence (c12); (b) three conditions that were a subset of the Jabberwocky conditions from PDD and were selected to span the range of constituent lengths: a list of 12 unconnected nonwords (jab-c01), three 4-word Jabberwocky constituents (jab-c04), and a 12-word Jabberwocky sentence (jab-c12); and (c) two nonconstituent conditions—four 3-word nonconstituent chunks (nc03) and three 4-word nonconstituent chunks (nc04). Sample stimuli are shown in Figure 1.

Like the materials in Experiments 1 and 2, the materials in Experiment 3 implicitly contained sequences of contiguous chunks of varying length drawn from English sentences. However, unlike Experiments 1 and 2, these chunks were not required to (and generally did not) form syntactic constituents in their source contexts. Thus, Experiment 3 allows us to investigate the extent to which constituency is critical to the relationship between implicit

chunk length and the brain's response. The materials for Experiment 3 consisted of two sets, so as to span a large range of chunk lengths at a fine-grained level. Stimuli in Set 1 were 24 words long in total and fell into length conditions based on the divisors of 24 (i.e., c01, c02, c03, c04, c06, c08, and c12). Stimuli in Set 2 were 30 words long and fell into length conditions based on the divisors of 30 (i.e., c01, c02, c03, c05, c06, and c10).

Full details about the materials and stimulus design are given in Appendix 4. Quantitative analyses of the linguistic features of these materials are given in Appendix 5.

Critical Task: Procedure

The procedure was similar for the three experiments and followed PDD: Participants saw the stimuli presented one word/nonword at a time in the center of the screen in all caps with no punctuation at the rate of 300 msec per word/nonword. In all experiments, participants were simply instructed to read attentively, based on prior evidence (Fedorenko et al., 2010) that responses to sentences, Jabberwocky sentences, word lists, and nonword lists do not appear to be affected by the presence of a task. In Experiment 1, the 150 trials (thirty 12-word stimuli × 5 conditions) were distributed across five runs, so that each run contained six trials per condition. In addition, each run included 108 sec of fixation, for a total run duration of 216 sec (3 min 36 sec). In Experiment 2, the 330 trials (thirty 12-word stimuli × 11 conditions) were distributed across 10 runs, so that each run contained three trials per condition. In addition, each run included 121.2 sec of fixation, for a total run duration of 240 sec (4 min). In both experiments, the order of conditions and the distribution of fixation periods in each run were determined with the optseq2 algorithm (Dale, Fischl, & Sereno, 1999). Experiment 3 used the same presentation format as Experiments 1 and 2, which means that the Set 1 (24-word) trials lasted 7.2 sec, and Set 2 (30-word) trials lasted 9 sec. The 156 trials of Experiment 3 (twelve 24-word stimuli × 7 conditions plus twelve 30-word stimuli × 6 conditions) were distributed across six runs, with each run containing 26 trials (fourteen 24-word trials, and twelve 30-word trials), two trials of each of the 13 conditions. Fixation periods were distributed as follows: 8 sec at the beginning of the run, 5.4 sec after each trial, and 8.2 sec at the end of the run. Condition order varied across runs and participants, with the constraint that trials of the same condition did not appear in a row.

fMRI Data Acquisition, Preprocessing, and Modeling; Functional Localization; and Data Analysis

fMRI data acquisition, preprocessing, and modeling are described in Appendix 2. Participant-specific functional localization is described in Appendix 3. Contrasts used in analyses of responses to the critical tasks are defined in Appendix 6. Statistical methods are described in Appendix 7.

Motivation for Functionally Localizing the Language Network

Here, we briefly motivate our assumption that there exists a core "language network" and our approach to identify this network at the individual-participant level using a validated language "localizer" paradigm (for further discussion of the importance of participant-specific localization of functional areas in the brain, see Kanwisher, 2010; Saxe et al., 2006; for a discussion of this issue in the domain of language specifically, see Braga et al., 2020; Fedorenko et al., 2010, 2011).

There Exists an Integrated Language Network in the (Typical, Mature) Human Brain

Several lines of evidence converge to support the view that parts of frontal and temporal cortex (among possibly other cortical, subcortical, and cerebellar areas; Lipkin et al., 2022; Fedorenko et al., 2010) constitute an integrated network that is implicated in language processing (for a recent review, see Fedorenko et al., 2024). First, within individuals, fMRI responses are stable (i.e., highly topographically similar) across diverse reading- and listening-based localizer contrasts that target high-level language processing, including coarse contrasts (such as sentences vs. nonword lists / consonant strings / acoustically degraded speech; Malik-Moraleda et al., 2022; Scott et al., 2017; Fedorenko et al., 2010) and finer contrasts (such as sentences vs. word lists, or word lists vs. nonword lists; Fedorenko et al., 2010). Second, within individuals, fMRI responses to localizer contrasts are stable over time (Mahowald & Fedorenko, 2016; Fedorenko et al., 2010). Third, within individuals, voxels within this network show highly correlated activity with each other during naturalistic comprehension paradigms and much higher than with voxels outside the language network (Malik-Moraleda et al., 2022; Paunov et al., 2019; Blank et al., 2014); in fact, these correlations are so strong that the same network that is identified by language localizers can also be recovered from patterns of BOLD signal fluctuations during a task-free resting state paradigm (Braga et al., 2020). Fourth, the same network is found across diverse languages, both across speakers (Malik-Moraleda et al., 2022) and within bi/multilingual speakers (Malik-Moraleda et al., 2024).

Thus, despite the inherent complexity both of language itself and of its (undisputed) interactions with diverse perceptual, motor, cognitive, and affective functions, evidence supports the hypothesis that the network identified by language localizer tasks is a functionally meaningful unit of analysis in the brain whose existence is external to, for example, conceptual debates about the definition of "language." The question thus becomes less a matter of how to

define language to study it in the brain, and more a matter of what computations this network supports. As of this writing, several lines of evidence suggest that "language processing" is the best available construal of this network's function. First, the perceptual controls used in many localizer tasks (e.g., sentences vs. *nonwords* or intact vs. *degraded speech*) rule out a low-level perceptual function, given that this network responds differentially to perceptually similar signals as a function of the presence of linguistic meaning and/or structure. Similarly, during language production, this network responds more strongly when individuals produce meaningful and structured language stimuli (phrases and sentences) compared with stimuli where the articulation demands are similar but the higher level language processing demands are not (Hu, et al., 2023), which suggests that low-level motor planning and execution cannot explain its responses. Second, this network is highly selective for language processing relative to diverse nonlinguistic inputs and tasks as measured with fMRI (Diachek, Blank, Siegelman, Affourtit, & Fedorenko, 2020; Deen, Koldewyn, Kanwisher, & Saxe, 2015; Monti, Parsons, & Osherson, 2012; Fedorenko et al., 2011; see Fedorenko et al., 2024, for review) and its damage impairs language production and comprehension but leaves intact diverse higher-order cognitive functions (Fedorenko & Varley, 2016). Third, this network is engaged by multiple levels of linguistic representation, from sublexical (Regev et al., 2024; Lopopolo et al., 2017), to lexical (Fedorenko et al., 2020; Rodd et al., 2005), to sentential (Shain et al., 2022; Caucheteux et al., 2021; Reddy & Wehbe, 2021) properties.

Therefore, the evidence for the existence of this integrated network is robust to many localizer-paradigm details and external to researcher assumptions about the definition of language (e.g., the network can be identified from resting state data). Although future research will likely continue to refine the precise computations that this network carries out, current evidence suggests that little precision is lost by calling this network the "language network," as we have done.

The Precise Locations of the Language Areas Vary between Individual Brains

The language network's general topography is similar across individuals (e.g., falling consistently within inferior and middle frontal gyri and along the superior temporal sulcus and/or middle temporal gyrus), and its precise topography is stable within individuals over time. However, the exact locations, shapes, and sizes of language areas vary between individuals (Lipkin et al., 2022; Fedorenko et al., 2010, 2011). This variability poses a problem for group-level analyses that average individual maps voxel-wise and/or use group-level ROIs (such as those used by PDD), given that these analyses likely pool responses from voxels that are highly selective for language with responses from nearby voxels that are less so or even belong to distinct functional networks (see Fedorenko & Blank, 2020, for illustration of this issue in inferior frontal cortex), which reduces sensitivity and functional resolution and underestimates effect sizes (Nieto-Castañón & Fedorenko, 2012; Saxe et al., 2006). These issues can lead to failure to detect real effects (e.g., length effects in Jabberwocky conditions in anterior temporal areas). Thus, of relevance to the current study, although PDD's parcels likely cover many core language areas (Appendix 9), they are also likely less precise given that they do not take into account interindividual variability in the precise locations of language areas (Fedorenko et al., 2011). Follow-up analyses of our data (Appendix 10) indicate that effects are indeed much weaker using PDD's group-level ROIs than they are using participant-specific fROIs.

The Use of Functional Localization Is Unlikely to Bias Our Results against Finding Separable Lexical, Syntactic, and Semantic Functions

One possible concern about our analysis design (raised by an anonymous reviewer) is that our localizer contrast (*sentences > nonword lists*) may be biased toward finding overlap between lexical semantic, syntactic, and combinatorial semantic processing demands, given that sentences and nonword lists differ in all three of these dimensions, and thus the areas that show the largest difference between them may also be those in which these demands overlap. This concern is partially addressed by Appendix 10, where we show that some of PDD's key findings (especially their reported syntax-selectivity of IFG) still fail to replicate when using their group ROIs instead of our functional ROIs. Moreover, this objection additionally rests on three questionable assumptions.

The *first assumption* is that there exist areas that selectively respond only to some of the demands of interest (lexical semantic, syntactic, and/or combinatorial semantic). However, this possibility has already been extensively investigated using narrower localizer contrasts, including *word lists > nonword lists* (targeting lexical processing), *Jabberwocky > nonword lists* (targeting "pure" syntactic processing), and *sentences > word lists* (targeting syntactic and combinatorial semantic processing), and no evidence of any such areas has emerged (Blank et al., 2016; Fedorenko et al., 2010, 2011; see also Bautista & Wilson, 2016, for related evidence from a different approach). Instead, areas/sets of voxels that respond to any of these contrasts tend to also respond to the other contrasts (Fedorenko et al., 2010), even at a finer spatial scale as measured with human intracranial recordings (Fedorenko et al., 2016). In fact, one of the motivating goals behind the original four-condition Fedorenko and colleagues (2010) localizer design was to separate selectivity for different types of linguistic representation. After no evidence of such separation emerged in multiple studies, Fedorenko and colleagues began to simplify the localizer design to the sentences versus nonwords version used here.

The *second assumption* is that simultaneously manipulating the demands associated with lexical semantic, syntactic, and combinatorial semantic processing will produce a larger response in areas that support multiple of these functions compared with areas that support only one. However, even if there existed pure syntax-processing regions as suggested by PDD, it does not follow that their response to sentences would be smaller than that of mixed-function regions simply because they have narrower functional selectivity. Sentences are substantially more syntactically complex than nonword lists, and syntax-selective regions should strongly differentiate between them. In other words, although a large response to our localizer contrast does not entail syntax selectivity, syntax selectivity does entail a large response to our localizer contrast.

The *third assumption* is that the network identified by a language localizer is highly dependent on the particular contrast and that a different contrast (e.g., a more narrowly syntactic contrast) would select a substantially different set of voxels. If this assumption is false (i.e., if roughly same set of voxels is picked out by many different contrasts), then the content of the particular localizer contrast becomes irrelevant, as long as it reliably identifies the same brain network as other approaches. As argued earlier in this section, we believe this is precisely the case for the language network. To re-emphasize one compelling piece of evidence, Braga and colleagues (2020) showed that the activation map for the *sentences > nonword lists* contrast tightly corresponds to a network that emerges from voxel timecourse correlations during task-free resting state. In other words, the contrast used in our study is simply an efficient way of identifying the network of interest (based on a few minutes of task data) that would emerge anyway if we had ~1 hr or more of resting-state data per participant.

On the basis of the considerations above, we find it unlikely that participant-specific functional localization led to spurious findings of overlap in our study.

RESULTS

We revisit PDD's claims in three experiments. Experiment 1 seeks to replicate the finding of an overall increase in the BOLD response of language brain areas as a function of chunk length. Experiment 2 is a conceptual replication of PDD, including all of the original real-word conditions and a critical subset of the Jabberwocky conditions, as well as PDD's two additional "nonconstituent" conditions consisting of three- and four-word chunks that do not form valid syntactic constituents (e.g., *over the floodlit*; Figure 1B). Unlike PDD, in all experiments, we independently localize the language network in each participant and use a fully within-subject design. Experiment 3 more directly targets the centrality of constituency for obtaining the length effect (stronger responses to longer chunks) by presenting participants with 24-word and 30-word stimuli that are composed of chunks of varying length (taken from naturalistic texts), which overwhelmingly do not form constituents in their source sentences (86.5% of the time; because Experiment 3 was originally designed with a different research goal in mind, avoiding constituents entirely was not a consideration). For details about these experiments, see the Methods section. Results are visualized in Figure 3 (full significance testing details are given in Appendix 8).

Do the Language Regions Show Length Effects?

For the real-word conditions, all regions show the pattern reported by PDD: significantly increasing activation as a function of chunk length, including a smaller increase at larger lengths (e.g., c06 to c12) in all three experiments (Figure 3B–E).

Do Any Language Regions Behave Like "Syntactic Hubs," Showing Identical Responses to the Real-word and Jabberwocky Conditions?

No language region shows the pattern (reported by PDD for inferior frontal and posterior temporal areas) of visually indistinguishable patterns of response in the real-word and Jabberwocky conditions. Instead, all language regions' responses are modulated by lexicality, either in the overall response, in the slope of the length effect, or both (Figure 3C and E). Thus, no language region appears to be a hub for abstract (i.e., content-independent) combinatorics.

Do Anterior Temporal and Temporoparietal Language Regions Only Show Length Effects in the Real-word Conditions?

We find a significant length effect in the Jabberwocky conditions for the language network as a whole, as well as for each region within it except for the temporoparietal left angular gyrus (LAngG) language region. Contrary to PDD's claim that the anterior temporal language area (LAntTemp) is not responsive to chunk length in the absence of lexical content (Jabberwocky), we find this effect robustly (Figure 3C and E).

However, the LAngG language region (which corresponds to PDD's "TPJ" region; Figure A1) only shows a length effect in the real-word conditions (which is significantly larger than the length effect for Jabberwocky conditions), as PDD claimed, and in direct pairwise comparisons between regions, the length effect for Jabberwocky stimuli is significantly weaker in the LAngG functional ROIs (fROI) than in all other language regions. Nonetheless, the length effect for real-word stimuli is also significantly weaker in LAngG than in all other language regions except for the LAntTemp fROI. Together with prior evidence (e.g., Shain, Paunov, Chen, Lipkin, & Fedorenko, 2023; Shain, Blank, Fedorenko, Gibson, & Schuler, 2022; Braga et al., 2020; Blank, Kanwisher, & Fedorenko, 2014), this qualitative difference in response suggests functional differentiation

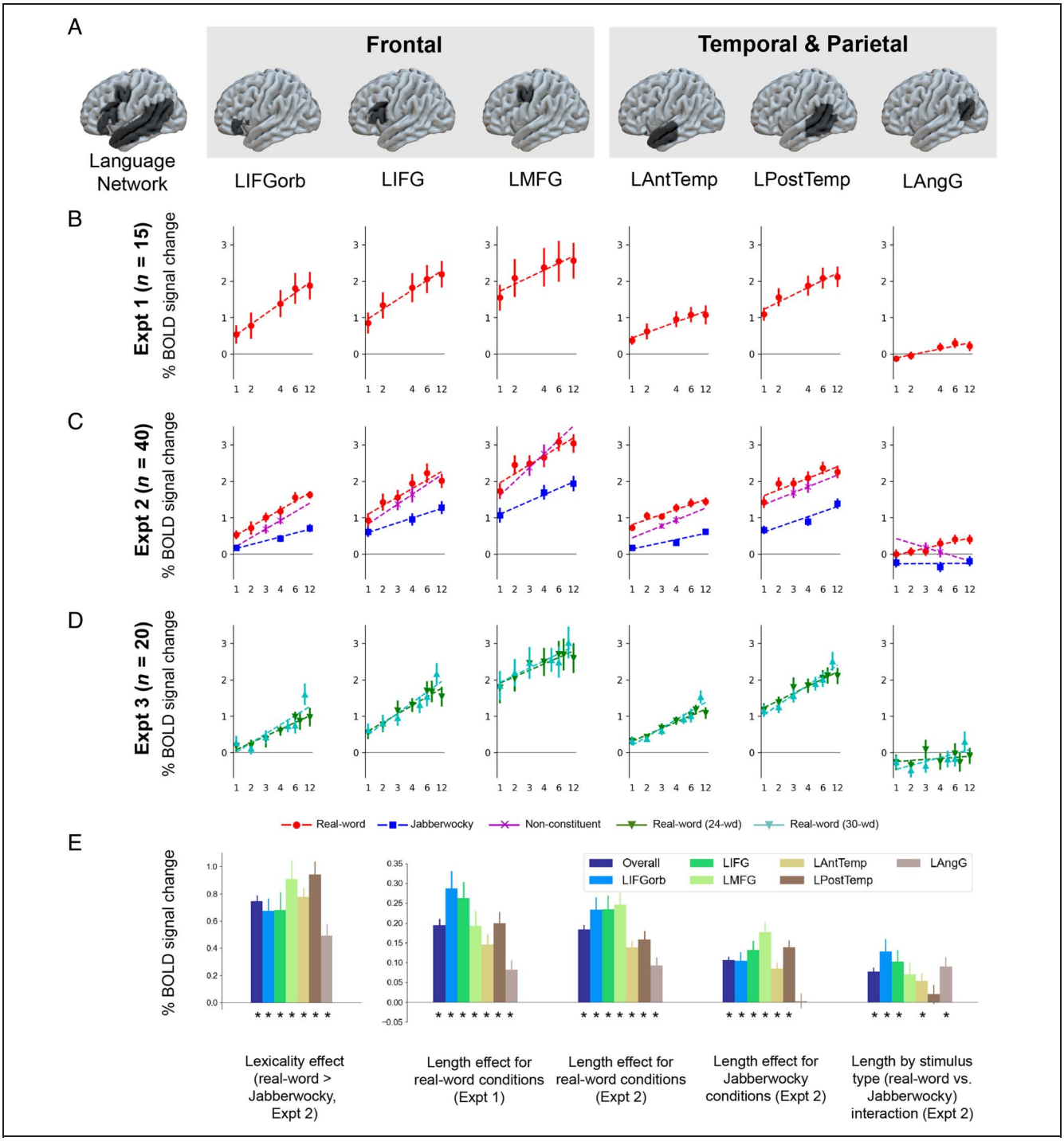

Figure 3. (A) Group masks bounding the six left-hemisphere regions of the language network. The top 10% of language-responsive voxels (i.e., voxels that respond to the localizer contrast, sentences > nonwords) are selected within each mask in each participant (see Methods section). (B) Estimated response to each real-word condition in Experiment 1 (which did not include Jabberwocky conditions). Responses in all regions increase with chunk length. (C) Estimated response to each real-word, Jabberwocky, and nonconstituent condition in Experiment 2. Responses in all regions increase with chunk length in the real-word conditions, and responses in all regions but the LAngG language region increase with chunk length in the Jabberwocky and nonconstituent conditions. (D) Estimated response to each condition of both the 24-word and 30-word items of Experiment 3, both of which consisted of contiguous real-word chunks that generally did not form syntactic constituents. Responses in all regions increase with chunk length to a similar degree as in the real-word conditions of Experiments 1 and 2. (E) Key contrasts by language network fROI (left-to-right): overall lexicality effect (increase in response to real-word over Jabberwocky conditions in Experiment 2, averaging over chunk length); length effect for real-word conditions in Experiment 1 (slope of the line by participant from B; length effect for real-word conditions in Experiment 2 (slope of the red line by participant from C); length effect for Jabberwocky conditions in Experiment 2 (slope of the blue line by participant from C); increase in length effect in real-word conditions over Jabberwocky in Experiment 2 (difference between the slopes of the red and blue lines by participant from C). Starred bars indicate statistically significant effects by likelihood ratio test (corrected for false discovery rate across fROIs; Benjamini & Yekutieli, 2001; see Appendix 8 for full testing results). Error bars show standard error of the mean over participants.

between the LAngG language region and the rest of the core language network (see Discussion section).

Are Inferior Frontal and Posterior Temporal Language Regions Insensitive to Combinatorial Semantics, over and above Syntax?

We find a significantly steeper slope for the length effect in the real-word conditions relative to the Jabberwocky conditions (Length × Stimulus Type interaction) in the language network as a whole, as well as in each region within it except for the left middle frontal gyrus (LMFG) and left posterior temporal (LPostTemp) language regions. The Length × Stimulus Type interaction in the LMFG region is positive and similar in magnitude to that of other regions, but it fails to reach significance. By contrast, the Length × Stimulus Type interaction in the LPostTemp region is numerically near zero. This finding is contrary to PDD's claim that the inferior frontal language areas (left inferior frontal gyrus- [LIFG] and its orbital part [LIFGorb]) are equally sensitive to chunk length in real-word and Jabberwocky conditions (Figure 3C and E). However, the LPostTemp language region shows highly similar length effects for real-word and Jabberwocky stimuli, as PDD claimed, and in direct comparisons, the difference in length effect between the real-word conditions and the Jabberwocky conditions is significantly weaker in the LPostTemp region relative to both the LIFGorb and the LIFG language regions, in spite of the fact that the left inferior frontal gyrus (IFG) has been classically associated with syntactic processing (e.g., Grodzinsky, Pieperhoff, & Thompson, 2021; Friederici, 2011; Hagoort, 2005; Caramazza & Zurif, 1976). This result supports PDD's claim that the LPostTemp region is insensitive to combinatorial semantics, showing equal sensitivity to syntactic structure, with or without lexical content. We return to this finding in the Discussion section.

Does Syntactic Constituency Critically Drive the Length Effect?

The length effect in Experiment 2 is at least as strong in the nonconstituent conditions as it is in the real-word constituent conditions, which undermines PDD's claim that length effects are driven primarily by syntactic constituency. This finding is reinforced by Experiment 3, which evaluates length effects in materials composed primarily (86.5%) of nonconstituents (Figure 3D). As shown, the length effect in response to these largely nonconstituent materials is qualitatively similar to the length effects reported in Experiments 1 and 2, and quantitatively, we observe no significant differences in any region, or in the language network as a whole, between the length effect in Experiment 3 versus in either Experiment 1 or Experiment 2 in between-groups comparisons. Thus, syntactic constituency does not critically drive the length effects in the language network.

Summary

Our results support a distributed burden of lexical-semantic, syntactic, and combinatorial-semantic processing throughout the language network (rather than the dissociation between syntactic and lexico-semantic subnetworks, as claimed by PDD) and challenge the claim that stronger responses to longer chunks are driven by syntactic constituency (given that these length effects are equally strong regardless of whether the chunks form constituents). The key similarities and differences between our findings and PDD's are summarized in Table 2.

Table 2. Summary of Key Similarities and Differences between PDD's Findings and Those of Our Study with Respect to Sensitivity to Lexical Content, Syntactic Structure, and Combinatorial Semantics (Differences Highlighted in Gray)

	Sensitivity to Lexical Content (Inconsistent with a Purely Syntactic Function)		*Sensitivity to Structure in Jabberwocky (Inconsistent with Purely Semantic Function)*		*Greater Sensitivity to Structure in the Presence of Lexical Content (Inconsistent with a Selectively Syntactic—vs. Combinatorial Semantic—Function)*	
	PDD	*This Work*	*PDD*	*This Work*	*PDD*	*This Work*
Inferior frontal	−	+	+	+	−	+
Anterior temporal	+	+	−	+	+	+
Posterior temporal	−	+	+	+	−	−
AngG/TPJ	+	+	−	−	+	+

PDD reported (a) one set of regions (inferior frontal and posterior temporal) that were sensitive to structure (chunk length) in real-word conditions and equally sensitive to structure in Jabberwocky conditions (supporting abstract syntactic processing in these regions), and (b) another set of regions (anterior temporal and TPJ) that were sensitive to lexical content and insensitive to structure in Jabberwocky conditions. Our study does not reproduce several of PDD's reported insensitivities (red minus signs) and challenges the purported double dissociation between semantic regions on the one hand (anterior temporal and temporoparietal areas) and syntactic regions on the other (inferior frontal and posterior temporal areas). Instead, we find more broadly distributed lexical semantic, syntactic, and combinatorial semantic effects throughout the language network, albeit with evidence (consistent with PDD's claims) that the temporoparietal area is only sensitive to structure in real-word conditions and that the posterior temporal language area is equally sensitive to structure in both real-word and Jabberwocky conditions.

See Appendix 9 for evidence that the results above hold when we use the masks from PDD to define the language areas, and see Appendix 10 for evidence that qualitatively similar patterns hold even when we average over PDD's entire ROI parcels (i.e., without functional localization), albeit with weaker overall effects. See Appendix 11 for evidence that the extremes of the length conditions—(jab-)c01 and (jab-)c12—replicate an established pattern of response in the language network. See Appendix 12 for exploratory analyses of the right-hemisphere homotopes of the left-hemisphere language areas.

DISCUSSION

Whether different brain areas specialize for different types of linguistic processing is a long-standing open question in the neurobiology of language. Perhaps the most frequently proposed pattern of specialization is a dissociation between some brain areas that selectively represent and process the syntax of sentences and others that selectively represent and process their semantics (Friederici, 2017; Dehaene et al., 2015; Bornkessel-Schlesewsky & Schlesewsky, 2013; Baggio & Hagoort, 2011). This perspective is inspired in part by the fact that some nonlinguistic domains (e.g., mathematics, action planning, and music) also exhibit a kind of "syntax" in that they obey similar principles to language of sequential, hierarchical, and symbolic representation (Lashley, 1951). If, as some have argued (Fitch & Martins, 2014; Koechlin & Jubault, 2006; Novick, Trueswell, & Thompson-Schill, 2005; Patel, 2003), abstract syntactic structure building is supported by a shared brain network with a key locus in the inferior frontal cortex, then the human capacity for language may be linked to a more general capacity for structured symbol manipulation, which may in turn have arisen from anatomical changes to pFC during human evolution (Dehaene et al., 2015, 2022). This position offers tantalizing continuities between language and other domains, along with explanatory links to evolutionary processes that might have set the stage (a) for the emergence of language or (b) for the increasing sophistication of human cognition following language evolution. However, the empirical literature that is used to support this position (from both neuroimaging and neuropsychology) has relied on analyses that average brains in a common space and assume that a given spatial coordinate implements the same function across individuals—an assumption that is known to be incorrect for the language system and to lead systematically both to (i) failure to discover functional selectivities that are present in individual brains and (ii) conflation of functions that are distinct in individual brains (Fedorenko et al., 2010, 2011; Fedorenko & Kanwisher, 2009; Saxe et al., 2006). This concern extends to the finding of distinct syntactic and lexico-semantic processing centers by PDD, whose results are additionally subject to concerns about (i) reliance on between-groups comparisons to substantiate the claim of abstract syntactic processing and (ii) using the same data to define the ROIs and to statistically examine their responses. Because PDD's results have informed much subsequent theorizing about the neural basis of language and the structure of mental representations for language (e.g., Bornkessel-Schlesewsky et al., 2015; Dehaene et al., 2015; Bolhuis et al., 2014; Fitch, 2014; Petkov & Jarvis, 2012) and because of a growing effort in the field to replicate influential findings (Gilmore et al., 2017; Yarkoni & Westfall, 2017; Open Science Collaboration, 2015; Makel & Plucker, 2014; Simons, 2014; Pashler & Wagenmakers, 2012), here, we revisit PDD's claims across three fMRI experiments that address these methodological concerns.

Our findings robustly replicate PDD's discovery of parametric sensitivity in the language areas to the amount of linguistic context (increasing activation for longer spans of coherent text), as well as their finding that this pattern continues to hold in several areas even when lexical content is removed. Not only do we find this pattern across multiple experiments and in a different language (English) than the originally used French, but the effects are statistically indistinguishable across multiple independent groups of participants, which suggests that PDD uncovered a stable population-level signature of language comprehension in the brain (Fedorenko et al., 2016). This signature constitutes compelling evidence both that the brain's response is modulated by linguistic complexity and that syntax contributes to this modulation independently of meaning. This finding from PDD (replicated here) is thus an important explanandum in any theory of the brain basis of language comprehension.

However, our findings do not accord with PDD's proposed division of labor within the language network, namely, a double dissociation between syntactic and lexico-semantic subnetworks. Instead, our results reveal a more distributed pattern of lexical-semantic, syntactic, and combinatorial-semantic processing than that proposed by PDD (key similarities and differences between our findings and PDD's are summarized in Table 2). First, our results challenge the notion of pure syntactic hubs (i.e., the claim that inferior frontal and posterior temporal language areas respond identically to syntactic complexity across real-word and Jabberwocky conditions). Instead, we find large and statistically significant increases in the language network's response, including in the inferior frontal and posterior temporal areas, to real-word relative to Jabberwocky stimuli. This finding aligns with several prior studies (fMRI: Fedorenko et al., 2010—see Figure A5 for a direct comparison of the overlapping subset of conditions, and also Bedny, Pascual-Leone, Dodell-Feder, Fedorenko, & Saxe, 2011; magnetoencephalography: Matchin, Brodbeck, Hammerly, & Lau, 2019; intracranial recordings: Fedorenko et al., 2016) and with growing evidence for strong integration between syntax and semantics in the representations and computations that underlie language processing across fields and approaches, from linguistic theory (e.g., Goldberg, 2005; Pollard & Sag, 1994; Jackendoff, 1990; Kaplan & Bresnan,

1982), to psycholinguistics (e.g., Schuler & Wheeler, 2014; Pylkkänen & McElree, 2006; Kamide, Scheepers, & Altmann, 2003; MacDonald, Pearlmutter, & Seidenberg, 1994), to computational linguistics (e.g., Oh & Schuler, 2021; Dyer, Kuncoro, Ballesteros, & Smith, 2016; Mikolov, Chen, Corrado, & Dean, 2013; Manning & Schütze, 1999), to cognitive neuroscience (e.g., Kauf, Tuckute, Levy, Andreas, & Fedorenko, 2024; Merlin & Toneva, 2022; Anderson et al., 2021; Caucheteux, Gramfort, & King, 2021; Reddy & Wehbe, 2021; Fedorenko et al., 2016, 2020; Bautista & Wilson, 2016; Blank, Balewski, Mahowald, & Fedorenko, 2016; Fedorenko, Nieto-Castañón, et al., 2012; Keller, Gunasekharan, Mayo, & Corley, 2009).

Second, our results challenge the claim that inferior frontal areas are insensitive to semantic (as opposed to purely syntactic) composition. Instead, we find larger increases in response to chunk length in the real-word compared with the Jabberwocky conditions in both the LIFG and LIFGorb language areas.

Third, our results challenge the claim that anterior temporal areas only process combinatorial structure in the presence of lexical meaning. Instead, we find significant increases in response to chunk length in the Jabberwocky conditions (see also Brennan et al., 2012; Fedorenko et al., 2010; Figure A5).

Our results thus suggest greater spatial overlap in the brain among lexical-semantic, syntactic, and combinatorial-semantic processing than suggested by PDD, at least at the level of the macroanatomical areas (e.g., inferior frontal vs. anterior temporal vs. posterior temporal components of the language network).

Our results bear on linguistic composition in the general sense in which we have used this term (i.e., unifying word-level syntactic or semantic representations into phrase-level representations; Table 1). This general sense of *composition* should be distinguished from the narrower sense of composition as a transparent derivation of meaning via rule application (e.g., Montague, 1970), as opposed to, for example, the opaque conventionalized meanings of some multiword expressions (e.g., idioms like *let the cat out of the bag*). Our study does not attempt to distinguish pathways to phrasal meaning and therefore does not bear directly on the degree to which the brain relies on rules, surface statistics, and/or prior knowledge of the world to derive meaning from language (although this question has received substantial attention in the literature, e.g., Baggio, 2021). Our present concern is instead to characterize the effect of phrasal structure in brain areas responsible for inferring phrase-level syntactic and semantic representations, irrespective of how they do so.

We are by no means the first to express skepticism about the existence of sharp macroanatomical boundaries between syntax and semantics in the brain. Direct critiques of this idea have been raised both by our own group (e.g., Fedorenko et al., 2020) and by others (e.g., Aliko, Wang, Small, & Skipper, 2023; Rodd, Longe, Randall, & Tyler, 2010; Rodd, Vitello, Woollams, & Adank, 2015; Skipper, 2015; Wilson & Saygın, 2004). Indeed, several existing studies already support a broad distribution of syntactic (Shain et al., 2022; Shain, Blank, van Schijndel, Schuler, & Fedorenko, 2020) and semantic (Tang, LeBel, Jain, & Huth, 2023; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016) processing during language comprehension, as well as gradient (rather than categorical) differences in selectivity for syntax and semantics throughout much of perisylvian language cortex (Caucheteux et al., 2021; Reddy & Wehbe, 2021). Much of this countervailing evidence is based on naturalistic story listening data (although cf., e.g., Bautista & Wilson, 2016). Our study shows that even under experimental interventions that are highly similar to those used by PDD to produce some of the clearest evidence for a spatial dissociation between syntactic and semantic processing, results are most consistent with a distributed burden of lexical-semantic, syntactic, and combinatorial-semantic processing.

In addition to finding more distributed effects of both syntax and semantics than originally reported by PDD, our results also challenge the centrality to these length effects of syntactic constituency. A *constituent* is a complete phrase in a hierarchical representation of the sentence's grammatical structure (Figure 1B). PDD used constituents in the main conditions of interest, and—based on a comparison with control conditions that used nonconstituents—argued that the effect of chunk length critically depended on constituency, as opposed to other kinds of syntactic and semantic relations that hold between words in contiguous spans of language. Using PDD's narrow contrast between three- and four-word chunks that do not form constituents, we find that the increase in brain activity from the three-word condition to the four-word condition is at least as large in the nonconstituent stimuli as it is in the constituent stimuli in all regions except the LAngG language region (see below for discussion of this region). Furthermore, in a separate experiment that explored a wider range of implicit chunk lengths and consisted overwhelmingly (> 86%) of nonconstituent chunks, we find qualitatively and quantitatively similar effects of chunk length to those found when using valid syntactic constituents, with no significant difference in the length effect in any region or in the language network as a whole. This result is therefore incompatible with PDD's claim that chunk-length effects are driven primarily by the memory demands associated with assembling phrasal constituents. Nonetheless these length effects plausibly derive from linguistic complexity more broadly construed, and indeed, we find that multiple independently motivated measures of linguistic processing demand correlate with chunk length (Figure A2). Our results simply argue for an interpretation of length effects as driven by (perhaps diverse features of) richer linguistic *contexts*, rather than by phrasal constituency specifically. Other studies are needed to elucidate what those features are (see, e.g., Heilbron, Armeni, Schoffelen, Hagoort, & de Lange, 2022; Shain et al., 2020, 2022;

Lopopolo, Frank, van den Bosch, & Willems, 2017; Lopopolo, van den Bosch, Petersson, & Willems, 2021; Brennan & Hale, 2019; Brennan et al., 2012; Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016; Willems, Frank, Nijhof, Hagoort, & van den Bosch, 2016; Henderson, Choi, Luke, & Desai, 2015).

An alternative conceptualization of the length effects observed in our study and in PDD draws on the framework of "proper" and "actual" domains of specialized information processing systems (Barrett & Kurzban, 2006; Sperber, 1994), whereby the system's degree of engagement with an input can be modulated by the degree of fit between a given input and the target domain for which the system is adapted. Given the highly combinatory and contextualized nature of natural language, several words of contiguous context may be necessary to identify a stimulus as "proper" to the language network. As a consequence, shorter length conditions may fail by degrees to fully engage language processing mechanisms in the first place, thereby attenuating overall activation in the language network (see also Tuckute et al., 2024). Temporal receptive windows (TRWs), that is, the length of the preceding context that affects the processing of the current input (Lerner, Honey, Silbert, & Hasson, 2011; Hasson, Yang, Vallines, Heeger, & Rubin, 2008), could potentially serve as a filter for identifying domains proper to the language network, and indeed prior evidence supports the existence of TRWs for language on the order of a few words (Regev et al., 2023; Blank & Fedorenko, 2020; Nelson et al., 2017; Fedorenko et al., 2016; Lerner et al., 2011). However, the causes of these patterns of temporal receptivity are unknown, and they could derive from more basic kinds of linguistic processing (e.g., the degree to which nearby words can be composed into a syntactic parse may serve as a cue to whether an input is proper to the language network). In the absence of a deeper causal understanding of TRWs in the language network, viewing length effects as reflecting the distinction between proper and actual domains is not mutually incompatible with the interpretation whereby length effects reflect linguistic processing complexity.

Despite the lack of dissociation between syntactic and lexico-semantic processing centers and the broader distribution of diverse aspects of linguistic processing within the language network, our findings support two key functional asymmetries that were posited by PDD:

(1) *The LAngG/LTPJ language region differs functionally from the rest of the LH language network.*

First, like PDD, we find that the temporoparietal (LAngG in our terminology, left TPJ [LTPJ] in PDD's) language region behaves differently from the rest of the language regions: The length effect for Jabberwocky stimuli is (i) not significant and (ii) significantly smaller than in all other language regions. Thus, the LAngG language region is indeed less responsive to chunk length than other language regions in the absence of lexical content.

This finding should be interpreted in the context of related evidence that speaks to the role of this temporoparietal area in language processing. Although this area was identified as a language area by PDD and included as part of the language network in early studies using the functional localization paradigm (given its stronger responses to sentences than lists of pseudowords; Pallier et al., 2011; Fedorenko et al., 2010), evidence has accumulated over the last decade that this area differs functionally from the rest of the language network. First, the LAngG/LTPJ language region shows systematically weaker correlations with other language regions during naturalistic cognition paradigms than those regions do with each other (Malik-Moraleda et al., 2022; Paunov, Blank, & Fedorenko, 2019; Blank et al., 2014). Furthermore, data-driven functional parcellation using dense individual-subject resting state data picks out the same core temporal and frontal areas examined here—but not the LAngG/LTPJ region—as an integrated network (Braga et al., 2020). Second, the LAngG/LTPJ language region shows substantially weaker evidence, compared with the other LH language regions, of core language processing operations like next-word prediction and syntactic structure building (Shain, Blank, et al., 2020, 2022; Blank et al., 2016). Third, the LAngG/LTPJ language region responds at least as strongly to pictures and videos of meaningful events as to sentences, and sometimes more strongly (Shain, Paunov, Chen, et al., 2023; Ivanova et al., 2021; Amit, Hoeflin, Hamzah, & Fedorenko, 2017), contra other language regions, which are selective for linguistic over pictorial inputs. In addition, this region often shows *below-baseline* responses during language conditions (e.g., both in this study and in PDD), which would be explained if this area is instead a node in the *default mode network* (Vincent et al., 2006; Greicius, Krasnow, Reiss, & Menon, 2003; Raichle et al., 2001), a brain network whose activity increases during rest, and which has been associated with high-level conceptual processing and/or episodic self-projection (Davey et al., 2016; Philippi, Tranel, Duff, & Rudrauf, 2015; Vincent et al., 2006; Greicius et al., 2003). Many have implicated the angular gyrus broadly (cf. the language-responsive portion of it) in heteromodal conceptual integration (Ivanova et al., 2021; Davis & Yee, 2019; Amit et al., 2017; Fernandino et al., 2016; Price, Bonner, Peelle, & Grossman, 2015; Price, Peelle, Bonner, Grossman, & Hamilton, 2016; Bonner, Peelle, Cook, & Grossman, 2013; Seghier, 2013; Binder, Desai, Graves, & Conant, 2009). This hypothesis could explain the greater response in the AngG/TPJ region to meaningful language stimuli, even in the absence of a selectively linguistic function. As a result of all this evidence, in recent work (e.g., Chen et al., 2023; Shain, Paunov, Chen, et al., 2023), we have begun excluding the LAngG language area from our definition of the language network.

(2) *The LPostTemp language region is sensitive to syntax and lexical semantics but not combinatorial semantics.*

The claim from PDD that our results most strongly support is that the language-responsive area in the posterior temporal cortex is similarly sensitive to syntax, with or without lexical content. Although the overall response of the LPostTemp region to real-word stimuli is greater than its response to Jabberwocky stimuli, the difference in the length effect between real-word and Jabberwocky stimuli is virtually zero, as evidenced by similar slopes (Figure 3C and E)—the +Lex, +Synt, –CombSem profile in Figure 2. This result is inconsistent with PDD's strong characterization of LPostTemp as a pure syntactic hub (−Lex, +Synt, –CombSem), given that its response is strongly influenced by lexical content, independently of syntax. However, it does suggest that the burden of combinatorial processing in the LPostTemp region is unaffected by the meaningfulness of the resulting structure, which supports a *lack of combinatorial-semantic processing over and above syntactic processing*. This profile appears to be unique to the LPostTemp language region; the difference between the length effects in real-word versus Jabberwocky stimuli is nonsignificant and near zero in LPostTemp, significant in inferior frontal (LIFG and LIFGorb) language regions, and significantly larger in the frontal regions than in the LPostTemp region in direct comparisons, although LIFG has been classically associated with abstract syntax (Grodzinsky et al., 2021; Friederici, 2011; Hagoort, 2005; Caramazza & Zurif, 1976).

This result is important for two reasons. First, it lends support to the hypothesis that the posterior temporal language area plays a special role in processing hierarchical syntax, relative to other language areas that frequently co-activate with it during language processing (Matchin & Hickok, 2020; Bornkessel-Schlesewsky & Schlesewsky, 2013). Second, it is to our knowledge the first clear evidence of region-level (cf. Regev et al., 2023) functional differentiation within the human language network using functional localization methods—which account for inter-individual variation in the precise locations of language areas—and appropriate Region × Condition interaction statistics (e.g., Nieuwenhuis, Forstmann, & Wagenmakers, 2011). These methods have so far yielded a highly distributed picture of linguistic (including sublexical, lexical semantic, syntactic, and combinatorial semantic) processing across the regions of the language network, with little evidence of network-internal structure (Regev et al., 2024; Shain et al., 2020, 2022; Mollica et al., 2020; Blank et al., 2016; Fedorenko et al., 2010; see Fedorenko et al., 2024, for review). Our current results support invariance in the LPostTemp language region to combinatorial semantics (over and above syntax, –CombSem in the terminology of Figure 2). This finding is noteworthy in light of evidence that damage to posterior temporal cortex is associated with more severe and longer-lasting aphasia compared with other parts of the language network (Wilson et al., 2023). This finding also aligns with prior proposals that posterior temporal cortex may serve a critical early stage of the comprehension process: receiving input from perceptual areas (e.g., speech perception areas (Overath, McDermott, Zarate, & Poeppel, 2015)), identifying grammatical categories and hierarchical phrasal relations, and relaying this syntactic information to downstream conceptual semantic areas (Matchin & Hickok, 2020). This hypothesis could account for the apparent absence of combinatory semantic effects in the LPost Temp region, given that this region may be upstream from these combinatorial semantic computations. However, invariance to combinatorial semantics is a weaker claim than the widespread interpretation of PDD as showing a selectively syntactic role for the posterior temporal language region (i.e., −Lex, +Synt, –CombSem; see Appendix 1).

Although we have improved on the methods and analyses used in some prior work on the neurobiology of natural language syntax and semantics, our study nonetheless has limitations. First, our findings of overlap between syntax and semantics pertain only to large-scale brain regions. We have focused on macroanatomy because this is the level at which most current neurobiological models posit functional dissociations within the language network (Friederici, 2017; Duffau et al., 2014; Bornkessel-Schlesewsky & Schlesewsky, 2013; Hagoort, 2005; Hickok & Poeppel, 2007). Of course, current results are compatible with the existence of functional differentiation at smaller spatial scales (within the regions that we have used as units of analyses, within voxels, or within neural populations / individual cells, as can be measured with intracranial recordings. (Regev et al., 2023). Second, our intended manipulation of the presence/absence of lexical semantics between our real-word and Jabberwocky materials may not be pure: Some semantic information may be inferred from pseudowords' form (Blasi et al., 2016) or context (Li, 1988), and syntactic information may be harder to recover in Jabberwocky sentences. Indeed, prominent theories of syntax assume that most syntactic information is stored alongside semantics in the mental lexicon (with only very abstract composition rules that assemble these syntactic fragments into larger structures), perhaps resulting in impoverished syntactic representations for pseudowords relative to real words (Goldberg, 2005; Steedman, 2001; Chomsky, 1995a; Pollard & Sag, 1994). If our Jabberwocky materials are simply harder to parse than our real-word materials (or than PDD's original Jabberwocky materials), this could explain the steeper length effects that we find in real-word versus Jabberwocky conditions. Although we cannot entirely rule out such a confound, worse parsing of Jabberwocky conditions is unlikely to be the primary explanation for our results: Syntactic structure is sufficiently available in our Jabberwocky materials to drive increases in processing demand throughout the language network, and even in one region (the LPostTemp language region) to the same extent as real-word materials. Third, our finding of chunk length effects leaves open a wide space of questions about the kinds of computations that drive these effects. Although

chunk length effects are consistent with PDD's assumption that chunk length indexes the size of the neural assembly needed to represent a parse tree for the chunk, other interpretations are possible (as discussed above with respect to TRWs).

In conclusion, we find lexicality effects in inferior frontal and posterior temporal language regions, length effects for Jabberwocky stimuli in the anterior temporal region (as well as all other language regions except the one in the angular gyrus), and length by lexicality interactions in the inferior frontal language regions (and other language regions except the posterior temporal region). This pattern of findings is summarized in Table 2. These results collectively support a broad distribution of sensitivity to syntax and semantics throughout the human language network, challenging PDD's hypothesized dissociation between language regions that selectively process abstract syntax and language regions that selectively process lexical and/or combinatorial semantics. Our results instead converge with growing evidence that linguistic representations and computations over a range of levels of description (phonological, lexical, syntactic, and combinatorial-semantic) are largely distributed across the language network (Regev et al., 2024; Shain et al., 2022; Blank & Fedorenko, 2020; Fedorenko et al., 2020; Bautista & Wilson, 2016; Fedorenko, Nieto-Castañon, & Kanwisher, 2012). We do find evidence of one key invariance argued for by PDD: Although the posterior temporal language region is more responsive to materials with lexical content, it shows no increase in response to combinatorial semantics over and above syntax. This finding deserves further investigation, including with more temporally sensitive methods, to ask whether this brain region may support an earlier stage of comprehension that focuses on identifying the words and the grammatical relations among them, with inferences about, for example, logical semantics (entities, relations, quantifiers, entailments, etc.) subsequently taking place in other language areas. However, our results show that the burden of lexical-semantic, syntactic, and combinatorial-semantic processing is distributed across diverse cortical areas, and that no single area or set of areas constitutes the syntax hub claimed by PDD and related work.

APPENDIX

Appendix 1: Extended Discussion of the Impact of Pallier and colleagues (2011)

Here, we provide an extended discussion of how the research community has tended to interpret PDD with respect to the neurobiological bases of syntactic versus lexico-semantic processing.

PDD's finding of virtually identical parametric increases in inferior frontal and posterior temporal language areas' activation with chunk length across both real-word and Jabberwocky stimuli strongly suggests that these regions comprise an autonomous syntactic "module" (the –Lex, +Syn, –CombSem profile in the terminology of Figure 2 of the main article; Fodor, 1983). We believe this is the most straightforward interpretation of PDD's emphasis on "the **relative independence** of syntax from lexico-semantic features" (p. 2526, emphasis ours). Subsequent work by the authors has been more explicit about this interpretation: "Remarkably, when the stimuli were 'delexicalized' by substituting all content words with meaningless pseudowords while maintaining all grammatical words and inflections, a core set of areas in left IFG and pSTS **continued to respond identically**, suggesting their central role in the construction of abstract syntactic trees" (Dehaene et al., 2015, p. 12, emphasis ours; see also Dehaene, 2019). This interpretation of PDD has been explicit in some studies (Kempen, 2014) and is at least implied by other studies citing PDD in support of a "modular" (Hertrich et al., 2016), "core" (Dehaene, Al Roumi, Lakretz, Planton, & Sablé-Meyer, 2022; Dehaene, 2019; Nelson et al., 2017; Pattamadilok et al., 2016; Wang et al., 2015), or "pure" (Hage & Nieder, 2016) syntax network. We therefore believe that an important component of PDD's influence has been the suggestion of an autonomous module or network for syntactic tree building that is insensitive to the content (meaning) of those trees and that therefore responds identically to both real and Jabberwocky constituents.

This strong position is difficult to sustain in the face of abundant evidence that inferior frontal and posterior temporal language areas respond more to real-word than Jabberwocky stimuli (+Lex in Figure 2, e.g., Matchin et al., 2017; Mahowald & Fedorenko, 2016; Fedorenko et al., 2010; Humphries et al., 2006; Mazoyer et al., 1993, inter alia). However, a weaker interpretation of PDD's inferior frontal and posterior temporal results focuses only on the absence of a difference in the *slope* of the parametric effect of constituent length between real-word and Jabberwocky stimuli, while allowing for a difference in overall response between the two condition types (the +Lex, +Syn, –CombSem profile in Figure 2). This position abandons the notion that these regions constitute an independent syntactic module (in PDD's words, the "independence of syntax from lexico-semantic features"), given that they are allowed to be sensitive not only to the demands of processing syntactic structures but also to the demands associated with processing real (but not Jabberwocky) words (e.g., retrieving and representing lexical meanings). Under this view, the key invariance in these regions that is supported by PDD's results is to *combinatorial-semantic content*, given that the increase in activation with syntactic complexity is not greater in the real-word conditions (which have combinatorial-semantic meaning) versus the Jabberwocky conditions (which arguably do not). Studies that cite PDD in favor of syntax selectivity in these regions must at minimum have this weak interpretation in mind (e.g., Nelson et al., 2017; Hage & Nieder, 2016; Hertrich et al., 2016; Fitch, 2014; Fitch & Martins, 2014;

Berwick, Beckers, Okanoya, & Bolhuis, 2012; Cappa, 2012), although the distinction between the weak and strong claims above is rarely made explicit.

In addition, PDD's finding of a length effect in anterior temporal and temporoparietal language areas only in the real-word (but not the Jabberwocky) conditions has been taken to support a selectively semantic function for these areas (the +Lex, –Syn, +CombSem profile in Figure 2), by PDD themselves and by work building on their results (Frankland & Greene, 2020; Zaccarella et al., 2017; Bautista & Wilson, 2016; Friston & Buzsáki, 2016; Skeide et al., 2016; Bornkessel-Schlesewsky et al., 2015; Zaccarella & Friederici, 2015; Wilson et al., 2014; Bornkessel-Schlesewsky & Schlesewsky, 2013).

Appendix 2: Data Acquisition, Preprocessing, and First-level Modeling

Data Acquisition

All data were collected at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. Whole-brain structural and functional data were collected using one of two configurations. Data from participants in Experiment 1 and Experiment 2 that were scanned before 2021 (n = 40) were acquired on whole-body 3 Tesla Siemens Trio scanner with a 32-channel head coil. For these participants, T1-weighted, magnetization prepared rapid gradient echo structural images were collected in 176 sagittal slices with 1-mm isotropic voxels (repetition time [TR] = 2530 msec, echo time [TE] = 3.48 msec, inversion time [TI] = 900 msec, flip = 8°). Functional, BOLD data were acquired using an EPI sequence with a 90° flip angle and using generalized autocalibrating partially parallel acquisitions (GRAPPA) with an acceleration factor of 2, with the following parameters: thirty-three 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, field of view in the phase encoding (A > > P) direction 200 mm and matrix size 96 × 96, TR = 2000 msec and TE = 30 msec. The first 10 sec of each run were excluded to allow for steady-state magnetization.

Data from participants in Experiment 2 and Experiment 3 that were scanned in 2021 or later (n = 35) were collected on a whole-body 3 Tesla Siemens PRISMA scanner with a 32-channel head coil, also at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. For these participants, T1-weighted, magnetization prepared rapid gradient echo structural images were collected in 208 sagittal slices with 1-mm isotropic voxels (TR = 1800 msec, TE = 2.37 msec, TI = 900 msec, flip = 8°). Functional, BOLD data were acquired using a simultaneous multi-slice EPI sequence with a 90° flip angle and using a slice acceleration factor of 2, with the following acquisition parameters: fifty-two 2-mm thick near-axial slices acquired in the interleaved order (with 10% distance factor), 2 mm × 2 mm in-plane resolution, field of view in the phase encoding (A > > P) direction 208 mm and matrix size 104 × 104, TR = 2,000 msec, TE = 30 msec, and partial Fourier of 7/8. For both functional sequences, the first 10 sec of each run were excluded to allow for steady-state magnetization.

Preprocessing

fMRI data were analyzed using SPM12 (release 7487), CONN EvLab module (release 19b) and other custom MATLAB scripts. Each participant's functional and structural data were converted from DICOM to NIFTI format. All functional scans were coregistered and resampled using B-spline interpolation to the first scan of the first session (Friston et al., 1995). Potential outlier scans were identified from the resulting subject-motion estimates as well as from BOLD signal indicators using default thresholds in CONN preprocessing pipeline (5 *SD*s above the mean in global BOLD signal change, or framewise displacement values above 0.9 mm, Nieto-Castanon, 2020). Functional and structural data were independently normalized into a common space (the Montreal Neurological Institute template; IXI549Space) using SPM12 unified segmentation and normalization procedure (Ashburner & Friston, 2005) with a reference functional image computed as the mean functional data after realignment across all timepoints omitting outlier scans. The output data were resampled to a common bounding box between Montreal Neurological Institute-space coordinates (−90, −126, −72) and (90, 90, 108), using 2-mm isotropic voxels and fourth order spline interpolation for the functional data, and 1-mm isotropic voxels and trilinear interpolation for the structural data. Last, the functional data were then smoothed spatially using spatial convolution with a 4-mm FWHM Gaussian kernel.

First-level Modeling

For both the language localizer task and the critical task, effects were estimated using a general linear model in which each experimental condition was modeled with a boxcar function convolved with the canonical hemodynamic response function (fixation was modeled implicitly, such that all timepoints that did not correspond to one of the conditions were assumed to correspond to a fixation period). Temporal autocorrelations in the BOLD signal timeseries were accounted for by a combination of high-pass filtering with a 128-sec cutoff, and whitening using an AR(0.2) model (first-order autoregressive model linearized around the coefficient α = .2) to approximate the observed covariance of the functional data in the context of restricted maximum likelihood estimation. In addition to main condition effects, other model parameters in the general linear model design included first-order temporal derivatives for each condition (included to model variability in the hemodynamic response function delays), as well as nuisance regressors controlling for the effect of slow

linear drifts, subject-motion parameters, and potential outlier scans on the BOLD signal. Resulting effect estimates reflect percent BOLD signal change (PSC).

Appendix 3: Participant-specific Functional Localization

Procedure

Sixty-five participants (out of 75 total) performed the localizer task in the same session as the critical task, and the remaining participants performed the localizer in a different session (for evidence that localizer activations are stable across scanning sessions, see Lipkin et al., 2022; Braga et al., 2020; Mahowald & Fedorenko, 2016). Most participants completed one or two additional tasks for unrelated studies. The entire scanning session lasted approximately 2 hr.

Localizer Task

The task used to localize the language network is described in detail in Fedorenko and colleagues (2010). Briefly, we used a reading task that contrasted sentences and lists of unconnected, pronounceable nonwords in a standard blocked design with a counterbalanced order across runs. This contrast targets higher-level aspects of language including, critically, lexical-semantic, syntactic, and compositional-semantic processing, to the exclusion of perceptual (speech or reading-related) and articulatory processes (see, e.g., Fedorenko & Thompson-Schill, 2014, for discussion). Stimuli were presented one word/nonword at a time. Participants were asked to read the materials attentively and to press a button at the end of each trial (included to help participants remain alert). Importantly, this localizer has been shown to generalize across different versions: the sentences > nonwords contrast, and similar contrasts between language and a degraded control condition, robustly activates the fronto-temporal language network regardless of the task, materials, modality of presentation, and particular language (Chen et al., 2023; Malik-Moraleda et al., 2022; Ivanova et al., 2020; Scott et al., 2017; Fedorenko et al., 2010). This includes generalization to both narrower contrasts (e.g., sentences > lists of unconnected words; Blank et al., 2016; Fedorenko et al., 2010) and broader contrasts (e.g., listening to passages > listening to acoustically degraded passages; in fact, this latter, auditory version of the localizer was used for two participants in Experiment 3 because of poor data quality in the visual language localizer, as described below; e.g., Malik-Moraleda et al., 2022; Scott et al., 2017). Furthermore, the same network robustly emerges from naturalistic-cognition paradigms (e.g., resting state, listening to stories, watching movies) using the data-driven functional correlation approach (Braga et al., 2020, see also Branco, Seixas, & Castro, 2020; Blank et al., 2014; Tie et al., 2014), suggesting that this network constitutes a natural kind in the brain, and our localizer contrast is simply a quick and efficient way to identify this network as needed for testing critical hypotheses about it.

The whole-brain maps for the language localizer are available at: https://osf.io/fduve/.

Definition and Validation of Language-responsive Functional Regions of Interest

For each participant (in each experiment), we defined a set of language-responsive fROIs using group-constrained, participant-specific localization (Fedorenko et al., 2010). In particular, each participant's map for the sentences > nonwords contrast from the language localizer task was intersected with a set of six binary masks (the maps for the language localizer are available at: https://osf.io/fduve/). These masks were derived from a probabilistic activation overlap map for the language localizer contrast in a large set of distinct participants (n = 220) using the watershed parcellation, as described in Fedorenko and colleagues (2010), and corresponded to relatively large areas within which most participants showed activity for the target contrast. These masks covered the fronto-temporal language network: three in the left frontal lobe falling within the IFG, its orbital portion, and the MFG, and three in the temporal and parietal cortex (Figure 3A of the main article). Within each mask, a participant-specific language fROI was defined as the top 10% of voxels with the highest t-values for the localizer contrast (see Lipkin et al., 2022 for evidence that the fROIs are similar when defined using a fixed statistical threshold).

For two participants, the data quality for the standard version of the language localizer was low; however, both had completed an alternative version of the localizer based on listening to short passages versus acoustically degraded versions of those passages (see Malik-Moraleda et al., 2022; Scott et al., 2017 for evidence that this version of the localizer identifies the same areas as the standard, reading-based localizer). For two additional participants, one run of the language localizer showed some fMRI artifacts; as a result, we used just one run for these participants (which is sufficient for identifying the language network).

Before examining the data from the critical experiments, we ensured that the language fROIs show the expected signature response (i.e., a stronger response to sentences than nonwords). To do so, we used an across-runs cross-validation procedure (e.g., Nieto-Castañón & Fedorenko, 2012), where one run of the localizer is used to define the fROIs, and the other run to estimate the responses, ensuring independence (e.g., Kriegeskorte et al., 2009). As expected, and replicating prior work (e.g., Blank et al., 2016; Mahowald & Fedorenko, 2016; Fedorenko et al., 2010, 2011, inter alia), the language fROIs showed a robust sentences > nonwords effect across the 73 participants with cross-validated localizer

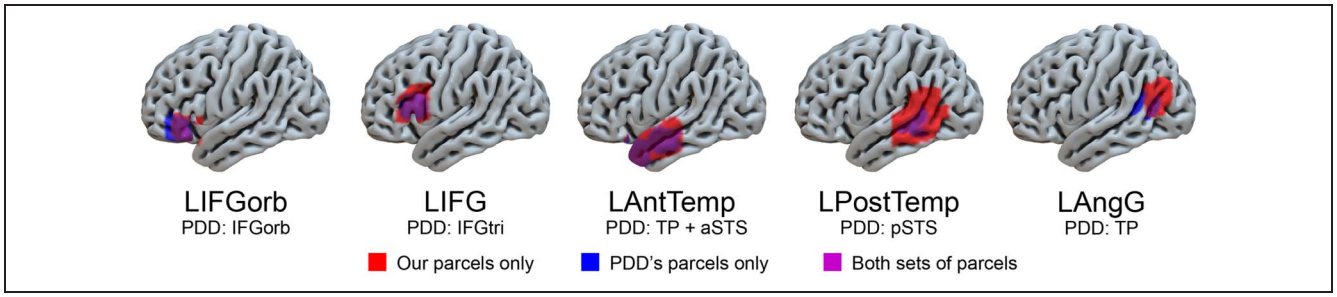

Figure A1. Visual comparison of language parcels used as group-level ROIs in PDD's original study versus to define individual-level fROIs in our study. Red voxels are only included in our parcels (but not PDD's), blue voxels are included in PDD's parcels (but not in ours), and purple voxels are included both in PDD's parcels and in our parcels. Overlap between the two sets of parcels is generally high.

contrast estimates (all $t_{(72)} > 7.26, p < 1e\text{-}9$, Cohen's $d > 0.085$), correcting for the number of regions (six) using the false discovery rate (FDR) correction (Benjamini & Yekutieli, 2001). The localizer activation maps for the two additional participants with a single run of the localizer task (preventing across-runs cross validation) were evaluated by visual inspection and looked typical.

Our masks show a close correspondence with the group-level ROIs used in PDD (Figure A1, see Appendix 9 for evidence that results replicate when using PDD's parcels as masks), with three exceptions: (i) PDD did not recover the language-responsive region in the MFG (because this region consistently emerges in both contrast-based and functional-correlation-based analyses as part of the language network—e.g., Braga et al., 2020; Glasser et al., 2016; Blank et al., 2014; Fedorenko et al., 2010—we chose to include it here); (ii) our AntTemp mask encompasses both of the anterior temporal ROIs in PDD (i.e., the anterior superior temporal sulcus ROI, and the temporal pole ROI; because PDD found similar functional profiles for these two ROIs, and for ease of comparisons with past work from our group, we chose not to split our mask into two parts); and (iii) our AngG mask only partially overlaps with PDD's TPJ mask (however, none of PDD's critical claims that we challenge in the current article pertain to this region; besides, our results for this region are similar to PDD's in spite of this difference in the masks, see Appendix 9).

Appendix 4: Materials Selection and Stimulus Design

Experiments 1 and 2

To create the materials for the real-word conditions, we extracted 180 two-word constituents (c02), 120 three-word constituents (c03), 90 four-word constituents (c04), 60 six-word constituents (c06), and 30 twelve-word constituents (c12) from the Penn-Treebank-parsed corpus (Marcus, Santorini, & Marcinkiewicz, 1993) and the Natural Stories corpus (Futrell et al., 2021). For each of the c02, c03, c04, and c06 conditions, the constituents were further manually concatenated into 30 twelve-word sequences, ensuring that syntactic or semantic dependencies would be unlikely to be formed across constituent boundaries. Finally, the c01 condition was created by selecting a set of 360 words from the full set of words in the Natural Stories corpus, and concatenating them into 30 twelve-word sequences, ensuring that adjacent words would be unlikely to combine syntactically or semantically.

To create the materials for the Jabberwocky conditions (jab-c01, jab-c04, and jab-c12), we took the strings from the c01, c04, and c12 real-word conditions and replaced all content words with pronounceable nonwords using the Wuggy software (Keuleers & Brysbaert, 2010).

To construct the materials for the nonconstituent conditions, we initially tried sampling three- and four-word nonconstituent spans from the Natural Stories corpus (Futrell et al., 2021), which contains hand-corrected phrase structure annotations. However, most strings extracted this way could often function as constituents in a different sentence context, especially given that many words in English can be used in multiple parts of speech. As a result, we hand-selected the nonconstituent chunks from a larger set of texts and manually concatenated them to ensure that syntactic or semantic dependencies were unlikely to be formed across boundaries. We used an online book recommendation app (available at recommendmeabook.com) to sample the first page of classic and recent best-selling fiction books (e.g., *The Poisonwood Bible* by Kingsolver). For every nonconstituent chunk, we extracted a nonconstituent of the appropriate length (three or four words long, depending on the condition) from a book and then manually searched for another nonconstituent that we believed would be unlikely to connect syntactically or semantically to the preceding one, and so on until the sequence (of four 3-word-long nonconstituents, or three 4-word-long nonconstituents) was complete. To protect against possible semantic dependencies, we often sampled nonconstituents from different books for the same sequence. Using this method, we created 30 twelve-word sequences for

each nonconstituent condition (out of 120 three-word nonconstituents for the nc03 condition, and out of 90 four-word nonconstituents for the nc04 condition). Sample stimuli from Experiments 1 and 2 are shown in Figure 1 of the main article, and the full set of materials is available on the Open Science Framework (https://osf.io/fduve/).

Experiment 3

To create the materials for the (largely nonconstituent) conditions of Experiment 3, we used the English Web Treebank of the Universal Dependencies corpus (Nivre et al., 2016). First, we removed sentences that consisted of fewer than 17 words (to permit variability in the starting position of a chunk within the source sentence, even for the longest, 12-word, chunks), which resulted in a treebank of 6273 sentences (out of the original 16,622). These sentences were randomly assigned to conditions, and one chunk of the appropriate length was then extracted from each sentence, starting from a randomly chosen word index (among Positions 1 through 5) within the sentence. We additionally required that: (i) no token in a chunk could be a proper noun, a punctuation mark, a number (containing any digits), or a symbol; (ii) no token in a chunk could have a miscellaneous/non-identified part-of-speech tag; (iii) the first two characters of any word could not be capitalized (to avoid abbreviations); and (iv) the chunk could not already be in the set of extracted chunks. We oversampled the number of sequences needed for each condition by a factor of three, to allow for subsequent filtering. We filtered sequences to ensure that the sets of strings were matched across conditions (with a p value of .05 or higher for any given condition pair in an independent-samples t test) in terms of the following features: (a) the average starting index of the string; (b) the ratio of content to function words (where content words included nouns, verbs, adjectives, and adverbs); (c) the average unigram lexical frequency; and (d) the average word length (in letters). If any pair of conditions was not matched on one or more of these features, the "worst offender" chunks were removed, and the statistics were recomputed. This was repeated until all pairs of conditions were matched on all features. The resulting set of chunks was then manually examined to remove chunks that straddled clausal boundaries or contained potentially sensitive, offensive, or highly culturally specific content. We then performed the matching described above one more time on the set of approved chunks, to ensure that no biases were introduced by the content filtering.

Having selected the set of candidate chunks, we developed an algorithm to concatenate them into 24-word and 30-word items. Following PDD, the key desideratum was that the boundaries between adjacent strings within an item be clearly detectable. A long short-term memory language model was trained on the English Web Treebank from which the chunks were sampled. Using this model, for all chunks of a given length, each possible chunk pair combination was assigned a cost calculated as $\log\left(\frac{p(s1+s2)}{p(s1)p(s2)}\right)$, where "s1 + s2" denotes chunk concatenation, and is computed by the language model. These costs were accumulated in an adjacency matrix. The chunk order with the minimum cost was found by greedily solving an asymmetric travelling salesman problem to select a minimum cost path through the chunks. This procedure resulted in a set of concatenated chunks to be used in the experiments. To create the condition with chunks of Length 1, we used the words from the condition of chunk Length 2 because the chunk Length 1 condition should be most critically comparable to the next-length-up condition. However, because all conditions were well-matched for lexical properties, as described above, the words used in chunk Length 1 condition were automatically matched to all the other conditions, too.

Appendix 5: Linguistic Features

We analyzed the materials in our real-word conditions in Experiments 1–2 with respect to six linguistic features with independent empirical support (*open nodes, node closings, storage cost, integration cost, 5-gram surprisal*, and *probabilistic context-free grammar (PCFG) surprisal*; all measures elaborated below), to shed light on possible causes of the length effects originally reported by PDD and replicated in our study. Results are reported in Figure A2. As shown, many of these features are either positively or negatively correlated with constituent length in these materials, suggesting potential directions for research that attempts to ground these effects in theory-driven accounts of language processing. We expand on these findings below.

Measures Derived from Memory-based Accounts of Language Processing

Open nodes and node closings. Nelson and colleagues (2017) elaborated on PDD's proposal—in the context of a follow-up study that used intracranial recordings—by hypothesizing a parsing mechanism that consumes more and more working memory until the constituent ends, permitting a merge operation (Chomsky, 1995b) whereby the memory allocated to representing that constituent is released and neural activation drops proportionately. Thus, PDD's pattern of stronger activity for sequences made up of longer constituents is hypothesized to derive from an accumulation of working memory demand by the parser over the course of constituent processing, with higher average demand for longer contiguous spans of text, because they can contain longer constituents. In Nelson and colleagues (2017), these "build-up" processes were encoded in the measure *open nodes* (a form of *storage cost* associated with maintaining items in working memory), and processes associated with merge and

memory release were encoded in the measure *node closings* (a form of *integration cost* associated with retrieving and updating items in working memory).

We computed both of these measures as described in Nelson and colleagues (2017) from phrase structure trees in a generalized categorial grammar (Nguyen, van Schijndel, & Schuler, 2012) that were automatically generated for all stimuli using a probabilistic parser (van Schijndel, Exley, & Schuler, 2013) and hand-corrected by an expert annotator (parses and annotations available from the ModelBlocks repository: https://github.com/modelblocks/modelblocks-release). For these and other measures, each distinct constituent within an item was treated as independent by the model. For sequences made up of multiple constituents, the values were averaged across constituents to derive a single value for each sequence.

Although node closings has independent psycholinguistic support (e.g., Brennan et al., 2012, 2016; Hale, 2006), this predictor is anticorrelated with PDD's constituent-length manipulations and therefore cannot explain the effect (Figure A2, node closings). *Open nodes* is better correlated with PDD's expected pattern (Figure A2, open nodes).

Dependency locality theory storage and integration cost. The dependency locality theory (DLT; Gibson, 2000) is one of many theories of working memory use in human sentence processing (see also, e.g., Rasmussen & Schuler, 2018; Lewis & Vasishth, 2005; Gordon, Hendrick, & Johnson, 2001). It was selected for analysis based on evidence that it best characterizes activity in the language network among a range of existing memory-based theories (Shain et al., 2022). DLT effects have also been reported in behavioral studies (Chen, Gibson, & Wolf, 2005; Grodner & Gibson, 2005). The DLT posits measures that are conceptually related to the open nodes (storage) and node closings (integration) predictors discussed above. In the DLT, storage cost (Figure A2, storage cost) tracks the number of incomplete syntactic dependencies that must be maintained in memory. Integration cost (Figure A2, integration cost) tracks the difficulty of constructing a syntactic dependency as a function of the number of intervening discourse referents. The measures of integration cost that we use here incorporate modifications described in Shain, van Schijndel, Futrell, Gibson, and Schuler (2016) that discount the cost of preceding modifiers and coordinate structures and increase the cost of verbs, following theoretical and empirical support described in Shain and colleagues (2022). DLT measures were computed automatically from the hand-corrected phrase structure trees described above.

The storage cost measure is correlated with PDD's constituent-length manipulation; the integration cost measure is anticorrelated with PDD's manipulation and therefore cannot explain the effect.

Measures Derived from Surprisal-based Accounts of Language Processing

An alternative class of accounts of language comprehension with broad empirical support (e.g., Shain et al., 2020; Goodkind & Bicknell, 2018; Willems et al., 2015; Smith & Levy, 2013; Frank & Bod, 2011) focus on the predictability of incoming words in context (Levy, 2008; Hale, 2001). Here, we focus on two such measures, following Shain et al. (2020), who found support for both in neural

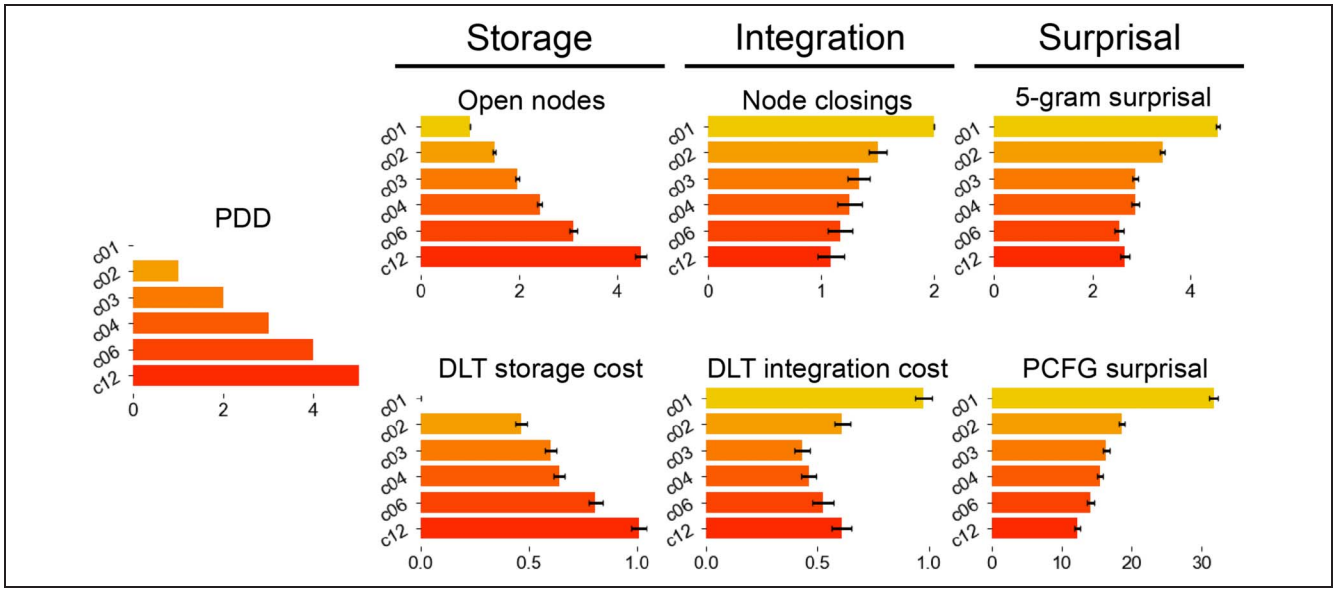

Figure A2. Mean value of linguistic features (memory- and surprisal-based) by constituent length for real-word conditions, compared with PDD-hypothesized monotonic increase (left). Error bars show standard errors of the mean across items.

responses of the language system during naturalistic story comprehension.

Five-gram surprisal. The negative log probability of a word in context as computed by KenLM 5-gram language models (Heafield, Pouzyrevsky, Clark, & Koehn, 2013) from frequency counts in the Gigaword 3 corpus (Graff, Kong, Chen, & Maeda, 2007; Figure A2, 5-gram surprisal). Five-gram models condition the probability distribution over the upcoming word on the sequence of four words that precede it, using default interpolation and backoff settings as described in Heafield and colleagues (2013). Five-gram models capture local word co-occurrence statistics but struggle to capture effects of larger-scale syntactic structures (e.g., constituency, long-distance dependencies). Robust effects of 5-gram predictability (and related models) are consistently reported in both behavioral (Smith & Levy, 2013; Frank & Bod, 2011; Demberg & Keller, 2008) and neuroimaging (Shain et al., 2020; Lopopolo et al., 2017; Willems et al., 2015) studies.

PCFG surprisal. The negative log probability of a word in context as computed by the PCFG parser of van Schijndel and colleagues (2013), trained on trees from the Penn Treebank (Marcus et al., 1993) that were automatically reannotated into a generalized categorial grammar formalism (Nguyen et al., 2012; Figure A2, PCFG surprisal). PCFG models condition only on hypothesized syntactic analyses of sentences. They therefore excel at capturing syntactic influences on expectations but struggle to capture local word-to-word cooccurrence patterns. PCFG effects have been reported in both behavioral (van Schijndel & Schuler, 2015; Fossum & Levy, 2012) and neuroimaging (Shain et al., 2020; Brennan et al., 2016) studies.

Both surprisal measures are anticorrelated with PDD's constituent-length manipulation and therefore cannot explain the effect.

Appendix 6: Contrast Definition for the Critical Experiments

The first-level models estimate the response in PSC to each condition of the critical experiment (e.g., c02, jab-c12). However, our critical research questions aggregate over these conditions in different ways (Is the response to real-word stimuli bigger than the response to Jabberwocky stimuli overall? Does activity increase on chunk length? etc.) Thus, as was done by PDD, we derive our key measures from the condition-level estimates. The resulting aggregate contrasts (estimated within each participant) are used as dependent variables for statistical analysis.

To estimate the overall response to real-word, Jabberwocky, and nonconstituent conditions, we computed a by-participant average of the responses to the stimuli in each of these broader stimulus types. To estimate the difference in response between real-word and Jabberwocky conditions, we took the by-participant difference between the averages within those two stimulus types only for Lengths 1, 4, and 12, which were represented for both stimulus types.

To estimate the parametric change in BOLD response as a function of constituent length, we computed the slope by participant of the best-fit line relating constituent length values to their associated first-level PSC estimates. To do so, we followed PDD in treating conditions c01, c02, c03, c04, c06, and c12 as equidistant, based on their observation of a sublinear monotonic relationship between length (in words) and the BOLD response. To model length effects in Experiment 3, which includes conditions not present in PDD's original study (i.e., Lengths 5, 8, and 10), we interpolated linearly between the points in PDD's original continuum. For example, Length 5 (which was not used by PDD) was treated as lying halfway between Lengths 4 and 6 (both of which were used by PDD). To estimate the difference in sensitivity to constituent length between stimulus types (e.g., between real-word and Jabberwocky conditions), we took the by-participant difference in slope between the two stimulus types.

Appendix 7: Statistical Analysis

We modeled the contrast values (as defined above, e.g., the by-participant difference in constituent length effect between real-word and Jabberwocky stimuli) as dependent variables in linear mixed-effects models in lme4 (Bates, Mächler, Bolker, & Walker, 2015) when examining entire networks, with random effects for participant and fROI, or simple linear models when examining the fROIs separately (since fROI-level models contain one contrast estimate per participant, there is no by-participant hierarchical structure to model). When examining the fROIs separately, reported p values are adjusted for false discovery rate (Benjamini & Yekutieli, 2001) over the number of fROIs in the network.

Network-wide contrast estimates were tested with the following mixed-effects model:

$$\texttt{Contrast} \sim \mathbf{1} + (1 \mid \texttt{Participant}) + (1 \mid \texttt{fROI}).$$

The critical variable in the above model is the intercept (**1**), which was tested by comparing this model (using a likelihood ratio test) to one in which the intercept is fixed at 0:

$$\texttt{Contrast} \sim \mathbf{0} + (1 \mid \texttt{Participant}) + (1 \mid \texttt{fROI}).$$

Regional contrast estimates were tested (against zero) using an unpaired t test. Pairwise tests of the difference in a contrast between two regions were tested in the same way, only using the difference in a given contrast from one region to the other (within an individual) as the dependent variable, rather than the contrast itself.

Appendix 8: Full Statistical Results from the Main Article

Full statistical results from the main article are reported in Table A1.

Table A1. Significance Tests of Key Contrasts with Estimates, Standard Errors, and *t* Values (Respectively Columns β, σ(β), and *t*)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
Constituent length for real-word conditions	1	**Overall**	0.19	0.04	5.34	< .001***
	1	**LIFGorb**	0.29	0.04	6.51	< .001***
	1	**LIFG**	0.26	0.04	6.39	< .001***
	1	**LMFG**	0.19	0.04	4.91	< .001***
	1	**LAntTemp**	0.15	0.03	5.62	< .001***
	1	**LPostTemp**	0.20	0.03	6.80	< .001***
	1	**LAngG**	0.08	0.03	3.28	.013*
Constituent length for real-word conditions	2	**Overall**	0.18	0.03	5.98	< .001***
	2	**LIFGorb**	0.23	0.03	7.37	< .001***
	2	**LIFG**	0.23	0.04	6.56	< .001***
	2	**LMFG**	0.25	0.03	8.20	< .001***
	2	**LAntTemp**	0.14	0.02	8.18	< .001***
	2	**LPostTemp**	0.16	0.02	7.72	< .001***
	2	**LAngG**	0.09	0.02	4.69	< .001***
Constituent length for Jabberwocky conditions	2	**Overall**	0.11	0.03	4.08	.003**
	2	**LIFGorb**	0.11	0.03	3.97	< .001***
	2	**LIFG**	0.10	0.02	4.78	< .001***
	2	**LMFG**	0.13	0.02	5.48	< .001***
	2	**LAntTemp**	0.18	0.03	6.97	< .001***
	2	**LPostTemp**	0.09	0.02	5.58	< .001***
	2	LAngG	0.14	0.02	8.30	1.000
Lexicality effect (real-word > Jabberwocky)	2	**Overall**	0.75	0.10	7.49	< .001***
	2	**LIFGorb**	0.67	0.09	7.43	< .001***
	2	**LIFG**	0.68	0.13	5.29	< .001***
	2	**LMFG**	0.91	0.14	6.68	< .001***
	2	**LAntTemp**	0.78	0.07	11.21	< .001***
	2	**LPostTemp**	0.94	0.09	9.94	< .001***
	2	**LAngG**	0.49	0.09	5.51	< .001***
Constituent-Length × Stimulus Type (real-word vs. Jabberwocky) interaction	2	**Overall**	0.08	0.02	3.28	.004**
	2	**LIFGorb**	0.13	0.03	4.03	.004**

Table A1. (*continued*)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
	2	**LIFG**	0.10	0.03	3.47	.006**
	2	LMFG	0.07	0.03	2.31	.078
	2	**LAntTemp**	0.05	0.02	2.63	.045*
	2	LPostTemp	0.02	0.02	0.84	1.000
	2	**LAngG**	0.09	0.02	3.68	.005**
Length effect in Experiment 3 (mostly nonconstituents)	3	**Overall**	0.19	0.03	5.72	< .001***
	3	**LIFGorb**	0.21	0.03	6.98	< .001***
	3	**LIFG**	0.26	0.04	6.82	< .001***
	3	**LMFG**	0.18	0.03	5.59	< .001***
	3	**LAntTemp**	0.20	0.02	9.87	< .001***
	3	**LPostTemp**	0.23	0.03	9.04	< .001***
	3	LAngG	0.06	0.03	2.42	.063
Length effect in Experiment 1 (constituents) vs. Experiment 3 (mostly nonconstituents)	1 & 3	Overall	0.00	0.03	−0.08	.938
	1 & 3	LIFGorb	−0.07	0.05	−1.44	1.000
	1 & 3	LIFG	0.00	0.06	−0.06	1.000
	1 & 3	LMFG	−0.01	0.05	−0.15	1.000
	1 & 3	LAntTemp	0.06	0.03	1.74	1.000
	1 & 3	LPostTemp	0.03	0.04	0.80	1.000
	1 & 3	LAngG	−0.02	0.04	−0.47	1.000
Length effect in Experiment 2 (constituents) vs. Experiment 3 (mostly nonconstituents)	2 & 3	Overall	0.01	0.03	0.25	.804
	2 & 3	LIFGorb	−0.02	0.05	−0.40	1.000
	2 & 3	LIFG	0.02	0.06	0.43	1.000
	2 & 3	LMFG	−0.06	0.05	−1.28	1.000
	2 & 3	LAntTemp	0.06	0.03	2.29	.316
	2 & 3	LPostTemp	0.07	0.03	2.07	.316
	2 & 3	LAngG	−0.03	0.03	−0.85	1.000
Difference in constituent length effect for real-word conditions	2	**LAngG vs. LIFGorb**	0.14	0.03	4.47	< .001***
	2	**LAngG vs. LIFG**	0.14	0.04	3.87	.002**
	2	**LAngG vs. LMFG**	0.15	0.03	5.27	< .001***
	2	LAngG vs. LAntTemp	0.05	0.02	1.91	.146
	2	**LAngG vs. LPostTemp**	0.07	0.02	3.15	.009**

Table A1. (*continued*)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
Difference in constituent length effect for Jabberwocky conditions	2	**LAngG vs. LIFGorb**	0.10	0.02	5.20	< .001***
	2	**LAngG vs. LIFG**	0.13	0.02	5.60	< .001***
	2	**LAngG vs. LMFG**	0.17	0.03	6.16	< .001***
	2	**LAngG vs. LAntTemp**	0.08	0.02	4.76	< .001***
	2	**LAngG vs. LPostTemp**	0.14	0.02	6.72	< .001***
Difference in Constituent-Length × Stimulus Type (real-word vs. Jabberwocky) interaction	2	**LPostTemp vs. LIFGorb**	0.11	0.03	4.30	.002**
	2	**LPostTemp vs. LIFG**	0.08	0.03	3.27	.021*

p Values are generated by likelihood ratio tests of linear mixed-effects models, with by-fROI results corrected for false discovery rate (FDR) applied over all six fROIs using the Benjamini–Yekutieli procedure (Benjamini & Yekutieli, 2001) using a nominal significance level of $\alpha = .05$. Starred *p* values indicate statistical significance under FDR correction (*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$). Significant regions are shown in **bold** in the fROI column.

Appendix 9: Results Replicate When Using PDD's ROIs as Masks to Define the Language fROIs

This study constrained the participant-specific functional localization procedure using broad masks for language areas that have been validated in prior work (e.g., Fedorenko et al., 2010). As discussed in Appendix 3, five out of six of these masks correspond closely to the ROIs reported in PDD, but the overlap between our masks and PDD's ROIs is not perfect. To ensure that our results are not because of the choice of the particular masks, in this section, we rerun our main analyses using PDD's language ROIs as localizer masks, rather than our standard localizer masks. As shown in Figure A3 and Table A2, results using PDD's parcels as localizer masks are highly similar to those reported using our standard masks in the main article, which indicates that results do not hinge critically on our choice of localizer masks.

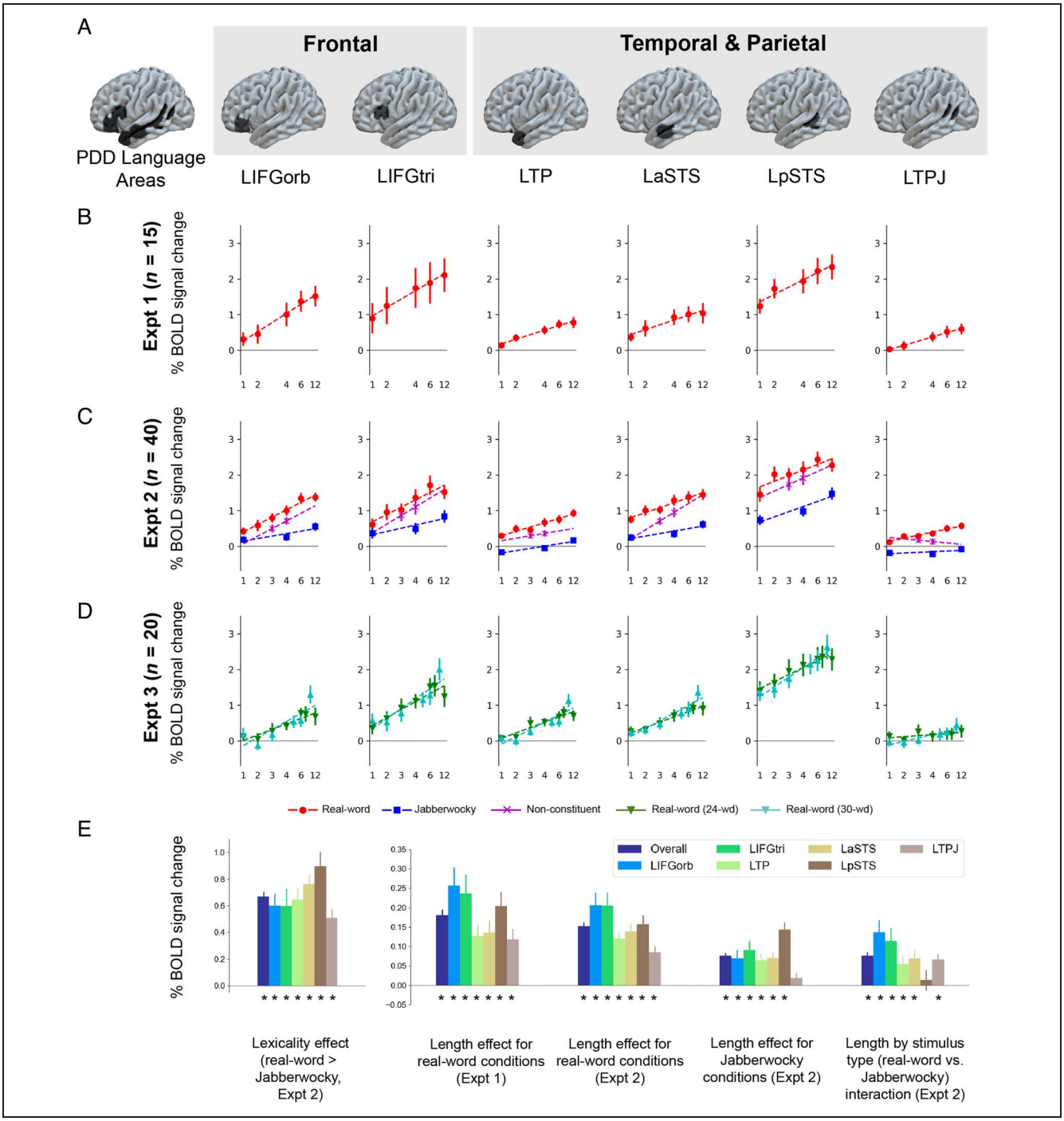

Figure A3. Main results (parallel of Figure 3 of the main article) using PDD's group-level ROIs as localizer masks, rather than the standard localizer masks. (A) PDD's group-level ROI parcels, used here as localizer masks. The top 10% of language-selective voxels are selected within each mask in each participant. (B) Estimated response to each condition of the real-word conditions in Experiment 1 (which did not include Jabberwocky conditions). Responses in all regions increase with constituent length. (C) Estimated response to each condition of the real-word conditions (replicating Experiment 1), the Jabberwocky conditions, and the nonconstituents conditions in Experiment 2. Responses in all regions increase with constituent length in the real-word conditions, and responses in all regions but LTPJ increase with constituent length in the Jabberwocky and nonconstituent conditions. (D) Estimated response to each condition of both the 24-word and 30-word items of Experiment 3, both of which consisted of contiguous real-word chunks that generally did not form syntactic constituents. Responses in all regions increase as a function of constituent length to a similar degree to the real-word conditions of Experiments 1 and 2. (E) Key contrasts by fROI (left-to-right): overall lexicality effect (increase in response for real-word over Jabberwocky conditions in Experiment 2, averaging over length); constituent-length effect for real-word conditions in Experiment 1 (slope of the line by participant from B); constituent-length effect for real-word conditions in Experiment 2 (slope of the red line by participant from C); constituent-length effect for Jabberwocky conditions in Experiment 2 (slope of the blue line by participant from C); increase in constituent-length effect in real-word conditions over Jabberwocky in Experiment 2 (difference between the slopes of the red and blue lines by participant from C). Starred bars indicate statistically significant effects by likelihood ratio test (corrected for false discovery rate across fROIs; Benjamini & Yekutieli, 2001). Error bars show standard error of the mean over participants.

Table A2. Reanalysis Using PDD's Group-level ROIs as Localizer Masks, Rather Than the Standard Language Localizer Masks (Parallels Table A1)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
Constituent length for real-word conditions	1	**Overall**	0.18	0.03	6.12	< .001***
	1	**LIFGorb**	0.26	0.05	5.48	< .001***
	1	**LIFGtri**	0.24	0.05	4.86	.001**
	1	**LTP**	0.13	0.03	4.64	.001**
	1	**LaSTS**	0.14	0.03	4.23	.002**
	1	**LpSTS**	0.20	0.04	5.52	< .001***
	1	**LTPJ**	0.12	0.03	4.26	.002**
Constituent length for real-word conditions	2	**Overall**	0.15	0.02	6.28	< .001***
	2	**LIFGorb**	0.21	0.03	6.44	< .001***
	2	**LIFGtri**	0.20	0.04	5.80	< .001***
	2	**LTP**	0.12	0.02	6.66	< .001***
	2	**LaSTS**	0.14	0.02	7.04	< .001***
	2	**LpSTS**	0.16	0.02	6.63	< .001***
	2	**LTPJ**	0.09	0.01	5.82	< .001***
Constituent length for Jabberwocky conditions	2	**Overall**	0.08	0.02	3.76	.003**
	2	**LIFGorb**	0.07	0.02	3.06	.012*
	2	**LIFGtri**	0.09	0.02	3.69	.003**
	2	**LTP**	0.06	0.02	3.96	.002**
	2	**LaSTS**	0.07	0.02	4.49	< .001***
	2	**LpSTS**	0.14	0.02	7.90	< .001***
	2	LTPJ	0.02	0.01	1.41	.389
Lexicality effect (real-word > Jabberwocky)	2	**Overall**	0.67	0.08	8.39	< .001***
	2	**LIFGorb**	0.60	0.09	6.73	< .001***
	2	**LIFGtri**	0.60	0.13	4.53	< .001***
	2	**LTP**	0.65	0.08	7.62	< .001***
	2	**LaSTS**	0.76	0.07	10.95	< .001***
	2	**LpSTS**	0.89	0.11	8.27	< .001***
	2	**LTPJ**	0.51	0.07	7.58	< .001***
Constituent-Length × Stimulus Type (real-word vs. Jabberwocky) interaction	2	**Overall**	0.08	0.02	3.34	.005**
	2	**LIFGorb**	0.14	0.03	4.44	< .001***
	2	**LIFGtri**	0.11	0.03	3.46	.006**
	2	**LTP**	0.06	0.02	2.67	.033*
	2	**LaSTS**	0.07	0.02	3.19	.010*
	2	LpSTS	0.01	0.03	0.51	1.000
	2	**LTPJ**	0.07	0.02	4.47	< .001***

Table A2. (*continued*)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
Length effect in Experiment 3 (mostly nonconstituents)	3	**Overall**	0.18	0.03	5.53	< .001***
	3	**LIFGorb**	0.19	0.03	7.35	< .001***
	3	**LIFGtri**	0.25	0.04	6.80	< .001***
	3	**LTP**	0.17	0.03	6.52	< .001***
	3	**LaSTS**	0.18	0.02	8.03	< .001***
	3	**LpSTS**	0.23	0.03	7.92	< .001***
	3	**LTPJ**	0.06	0.02	2.76	.030*
Length effect in Experiment 1 (constituents) vs. Experiment 3 (mostly nonconstituents)	1 & 3	Overall	0.03	0.03	0.92	.733
	1 & 3	LIFGorb	−0.02	0.05	−0.44	1.000
	1 & 3	LIFGtri	0.04	0.06	0.76	1.000
	1 & 3	LTP	0.05	0.03	1.70	.864
	1 & 3	LaSTS	0.04	0.03	1.22	.864
	1 & 3	LpSTS	0.07	0.04	1.82	1.000
	1 & 3	LTPJ	−0.03	0.03	−1.00	.864
Length effect in Experiment 2 (constituents) vs. Experiment 3 (mostly nonconstituents)	2 & 3	Overall	0.03	0.03	0.92	.362
	2 & 3	LIFGorb	−0.02	0.05	−0.44	1.000
	2 & 3	LIFGtri	0.04	0.06	0.76	1.000
	2 & 3	LTP	0.05	0.03	1.70	.697
	2 & 3	LaSTS	0.04	0.03	1.22	1.000
	2 & 3	LpSTS	0.07	0.04	1.82	.697
	2 & 3	LTPJ	−0.03	0.03	−1.00	1.000

Significance tests of key contrasts with estimates, standard errors, and t values (respectively columns β, σ(β), and t). p Values are generated by likelihood ratio tests of linear mixed-effects models, with by-fROI results corrected for false discovery rate (FDR) applied over all six fROIs using the Benjamini–Yekutieli procedure (Benjamini & Yekutieli, 2001) using a nominal significance level of $\alpha = .05$. Starred p values indicate statistical significance under FDR correction (*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$). Significant regions are shown in **bold** in the fROI column.

Appendix 10: Results Partially Replicate When Using PDD's ROIs as Group-level ROIs

We have advocated our use of participant-specific functional localization as a key methodological advantage of our study relative to PDD's (see Introduction and Methods sections of the main article). However, does this design choice impact results? Here, we investigate this question by following PDD's precedent and averaging responses across all voxels within each of PDD's parcels, without functional localization of participant-specific language areas. This approach thus uses the same set of voxels in all participants and does not account for interindividual variation in the precise locations of language areas.

Resulting estimates, plotted in Figure A4, are similar in important ways to those reported in the main article: In each region, responses increase (numerically) with chunk length in the real-word conditions as well as in the Jabberwocky conditions, albeit more weakly. However, as expected based on prior evidence (Fedorenko et al., 2010), sensitivity to all effects is greatly attenuated when using group-level ROIs as opposed to individual-level fROIs (Figure A4 and Table A3, see Figure A5 for the same visualizations with tighter y axes, for legibility); the only difference between the effects in Figure A4 and the comparatively stronger effects in Figure A3 is that the former averages over the entire parcel whereas the latter averages only over the 10% of the parcel that responds most strongly to the language localizer (as determined based on each individual map for the localizer contrast). This attenuation of effects is to be expected when group-level ROIs are used given that—for any given

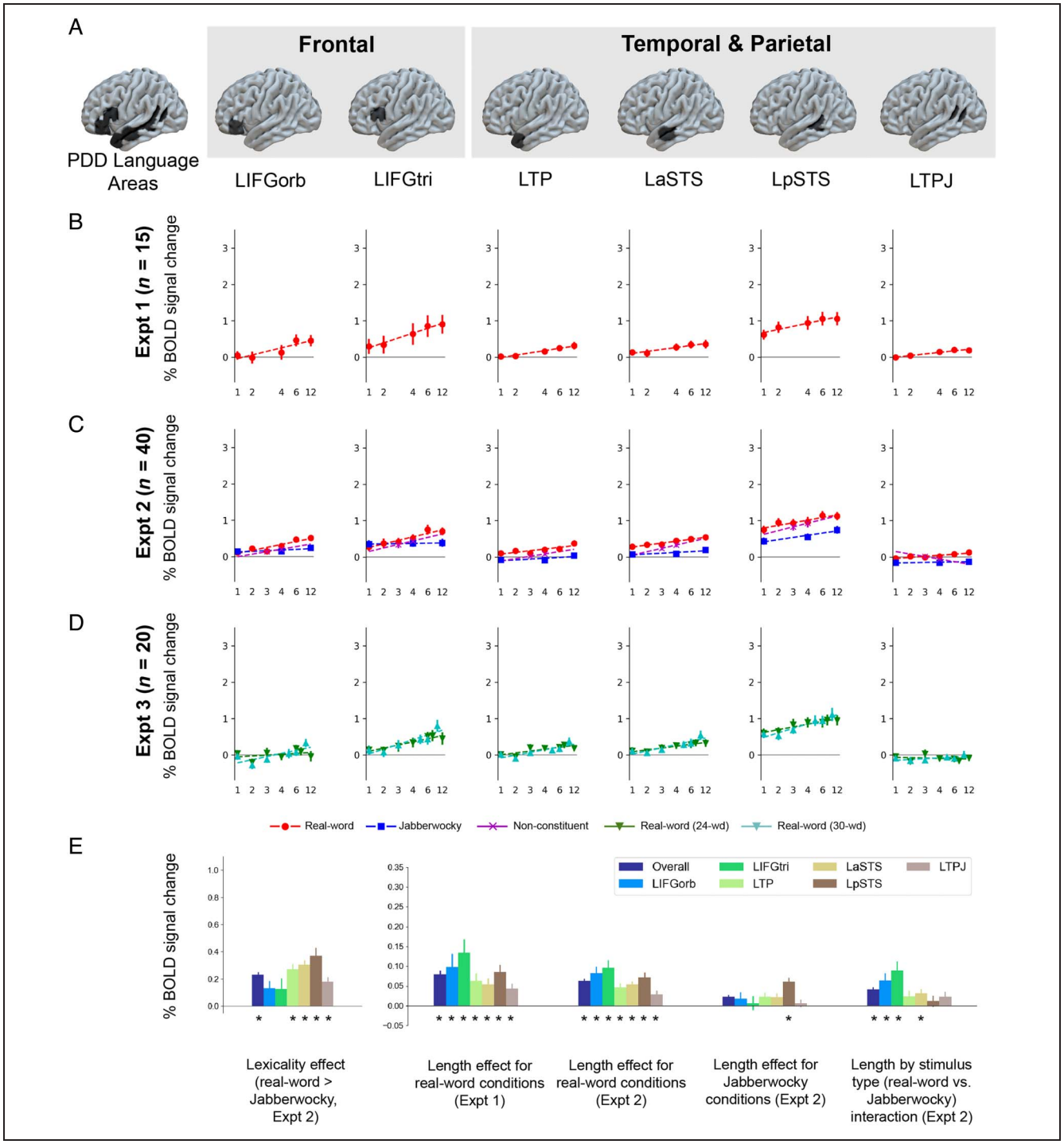

Figure A4. Main results (parallel of Figure 3 of the main article) using PDD's entire group-level ROIs, without functional localization. (A) PDD's group-level ROI parcels. Responses in all voxels of each parcel are averaged. (B) Estimated response to each condition of the real-word conditions in Experiment 1 (which did not include Jabberwocky conditions). Responses in all regions increase with constituent length. (C) Estimated response to each condition of the real-word conditions (replicating Experiment 1), the Jabberwocky conditions, and the nonconstituents conditions in Experiment 2. Responses in all regions increase with constituent length in the real-word conditions, but this increase is only significant in Jabberwocky conditions LpSTS. (D) Estimated response to each condition of both the 24-word and 30-word items of Experiment 3, both of which consisted of contiguous real-word chunks that generally did not form syntactic constituents. Responses in all regions but LTPJ increase as a function of constituent length, and to a similar degree to the real-word conditions of Experiments 1 and 2. (E) Key contrasts by fROI (left-to-right): overall lexicality effect (increase in response for real-word over Jabberwocky conditions in Experiment 2, averaging over length); constituent-length effect for real-word conditions in Experiment 1 (slope of the line by participant from B); constituent-length effect for real-word conditions in Experiment 2 (slope of the red line by participant from C); constituent-length effect for Jabberwocky conditions in Experiment 2 (slope of the blue line by participant from C); increase in constituent-length effect in real-word conditions over Jabberwocky in Experiment 2 (difference between the slopes of the red and blue lines by participant from C). Starred bars indicate statistically significant effects by likelihood ratio test (corrected for false discovery rate across fROIs; Benjamini & Yekutieli, 2001). Error bars show standard error of the mean over participants.

Table A3. Reanalysis of Average Responses in PDD's Group-level ROIs, without Functional Localization (Parallels Table A2)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
Constituent length for real-word conditions	1	**Overall**	0.08	0.02	4.70	< .001***
	1	**LIFGorb**	0.10	0.03	2.90	.028*
	1	**LIFGtri**	0.13	0.03	3.86	.013*
	1	**LTP**	0.06	0.02	3.23	.018*
	1	**LaSTS**	0.05	0.02	3.24	.018*
	1	**LpSTS**	0.09	0.02	4.79	.004**
	1	**LTPJ**	0.04	0.01	3.38	.018*
Constituent length for real-word conditions	2	**Overall**	0.06	0.01	4.92	< .001***
	2	**LIFGorb**	0.08	0.02	4.75	< .001***
	2	**LIFGtri**	0.10	0.02	4.97	< .001***
	2	**LTP**	0.05	0.01	4.43	< .001***
	2	**LaSTS**	0.05	0.01	6.88	< .001***
	2	**LpSTS**	0.07	0.01	5.71	< .001***
	2	**LTPJ**	0.03	0.01	2.92	.014*
Constituent length for Jabberwocky conditions	2	Overall	0.02	0.01	1.97	.061
	2	LIFGorb	0.02	0.02	1.10	1.000
	2	LIFGtri	0.01	0.02	0.37	1.000
	2	LTP	0.02	0.01	2.05	.229
	2	LaSTS	0.02	0.01	2.29	.200
	2	**LpSTS**	0.06	0.01	5.53	< .001***
	2	LTPJ	0.01	0.01	0.60	1.000
Lexicality effect (real-word > Jabberwocky)	2	**Overall**	0.23	0.05	4.70	< .001***
	2	LIFGorb	0.13	0.06	2.26	.086
	2	LIFGtri	0.13	0.08	1.60	.289
	2	**LTP**	0.27	0.04	6.70	< .001***
	2	**LaSTS**	0.31	0.03	8.98	< .001***
	2	**LpSTS**	0.37	0.06	6.19	< .001***
	2	**LTPJ**	0.18	0.03	5.12	< .001***
Constituent-Length × Stimulus Type (real-word vs. Jabberwocky) interaction	2	**Overall**	0.04	0.01	2.72	.017*
	2	**LIFGorb**	0.06	0.02	3.40	.012*
	2	**LIFGtri**	0.09	0.02	3.96	.005**
	2	LTP	0.02	0.01	1.60	.343
	2	**LaSTS**	0.03	0.01	2.92	.028*
	2	LpSTS	0.01	0.01	0.80	1.000
	2	LTPJ	0.02	0.01	1.91	.232

Table A3. (*continued*)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
Length effect in Experiment 3 (mostly nonconstituents)	3	**Overall**	0.06	0.02	3.96	.003**
	3	**LIFGorb**	0.05	0.01	4.03	.002**
	3	**LIFGtri**	0.10	0.02	4.68	< .001***
	3	**LTP**	0.06	0.02	4.14	.002**
	3	**LaSTS**	0.07	0.01	6.94	< .001***
	3	**LpSTS**	0.10	0.02	6.16	< .001***
	3	LTPJ	0.01	0.01	0.43	1.000
Length effect in Experiment 1 (constituents) vs. Experiment 3 (mostly nonconstituents)	1 & 3	Overall	−0.01	0.02	−0.88	.383
	1 & 3	LIFGorb	−0.05	0.03	−1.42	1.000
	1 & 3	LIFGtri	−0.03	0.04	−0.74	1.000
	1 & 3	LTP	0.00	0.02	0.01	1.000
	1 & 3	LaSTS	0.01	0.02	0.72	1.000
	1 & 3	LpSTS	0.01	0.02	0.49	1.000
	1 & 3	LTPJ	−0.04	0.02	−2.19	.525
Length effect in Experiment 2 (constituents) vs. Experiment 3 (mostly nonconstituents)	2 & 3	Overall	0.00	0.02	0.12	.908
	2 & 3	LIFGorb	−0.03	0.03	−1.16	1.000
	2 & 3	LIFGtri	0.01	0.03	0.28	1.000
	2 & 3	LTP	0.02	0.02	0.90	1.000
	2 & 3	LaSTS	0.01	0.01	1.05	1.000
	2 & 3	LpSTS	0.03	0.02	1.23	1.000
	2 & 3	LTPJ	−0.02	0.02	−1.47	1.000

participant—only a subset of the ROI may belong to the language network and the ROI may therefore include voxels that are not language-responsive. As a concrete example: When using fROIs, all regions (with the exception of the LAngG language fROI) show a length effect in the Jabberwocky conditions (Table A2), but when using group-level ROIs, only the LpSTS shows a length effect in Jabberwocky conditions (Table A3).

These reanalyses of our data using group-level ROIs differ from our main results (where individual-level functional ROIs are used) in the presence of several false negatives, which result from the lower sensitivity of group-level analyses (see also Fedorenko, Nieto-Castañón, et al., 2012; Nieto-Castañón & Fedorenko, 2012). These differences yield some outcomes that are more similar to those reported by PDD. In particular, using group-level ROIs, inferior frontal regions show no main effects of lexicality, and anterior temporal and temporoparietal regions show no significant Jabberwocky effects. Thus, differences between our main findings and those of PDD are plausibly due in part to differences in analysis methods, such that some of our findings emerge only when a more sensitive analytic approach is adopted, which takes interindividual variability in functional topography into account. However, the use of group-level ROIs does not fully explain the differences between our study and PDD's. For example, even when using the group-level ROIs (the same ROIs used by PDD), inferior frontal areas in our data show a significant length by lexicality interaction, such that responses increase more steeply with chunk length in real-word conditions compared with Jabberwocky conditions. This outcome is inconsistent with PDD's interpretation that inferior frontal areas belong to an abstract syntax network. Therefore, in addition to evidence of differences driven by analytic choices, we also find straightforward replication failures: A key finding reported by PDD does not appear to hold in our sample, even when the analyses are closely matched.

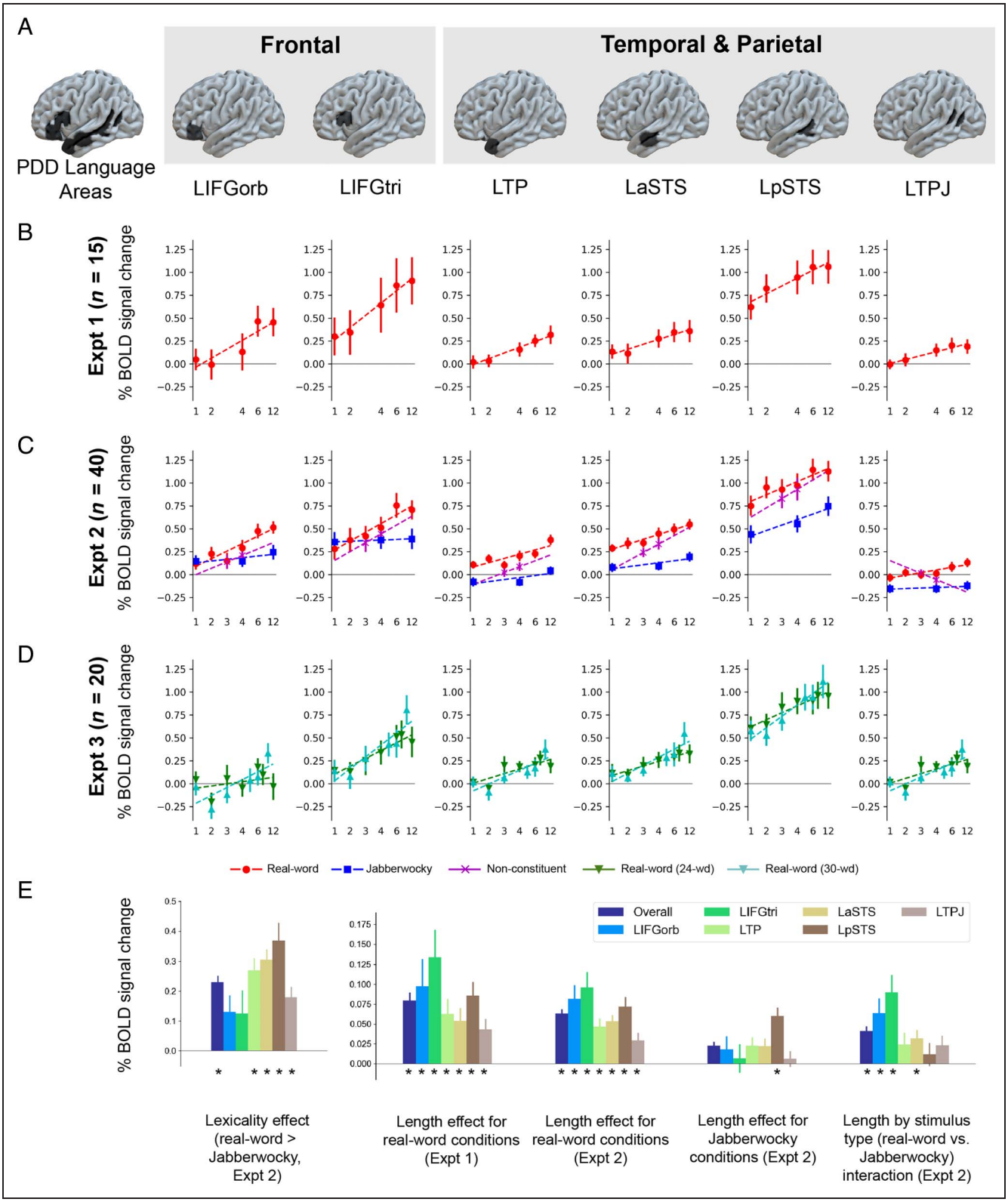

Figure A5. Identical to Figure A4, except with tighter *y* axis bounds for legibility.

Appendix 11: A Comparison of the Overlapping Sets of Conditions between Our Earlier Work (Fedorenko et al., 2010) and Experiment 2 in the Current Article

PDD conditions c12, c01, jab-c12, and jab-c01 correspond respectively to the sentence (S), word list (W), Jabberwocky (J), and nonword list (N) conditions that have been investigated in several prior studies, including by our group (Fedorenko et al., 2010). As shown in Figure A6, the pattern that we observed in Experiment 2 in the current study for this subset of conditions is remarkably similar to the patterns reported for Experiments 1 and 2 in Fedorenko and colleagues (2010); note that the difference in the overall response magnitude between the three experiments is most likely because of the fact that Experiment 1 in Fedorenko and colleagues (2010) and the current Experiment 2 used 12-word/nonword-long materials, and Experiment 2 in Fedorenko and colleagues (2010) used eight-word/nonword-long materials. Current Experiment 2 therefore constitutes a third within-laboratory replication—all with different sets of materials and non-overlapping sets of participants—of the pattern whereby sentences elicit the strongest response, word lists and Jabberwocky sentences intermediate response, and nonword lists the lowest response (see, e.g., Bedny et al., 2011, for another fMRI replication; see Fedorenko et al., 2016, for a replication in using electrocorticography). As discussed elsewhere, including in the main text, this pattern suggests that all the regions of the language network support both the processing of word meanings and combinatorial structure building.

Appendix 12: Analysis of Right Hemisphere Homotopic Regions

We have thus far followed PDD in exclusively analyzing LH language regions. In light of growing interest in the contribution of the right hemisphere (RH) to language processing (Martin et al., 2022), in this section, we include exploratory analyses of the key patterns within the RH homotopic language regions. Following, for example, Shain, Paunov, Chen, and colleagues (2023), we define these regions by first projecting the mirror images of our LH localizer masks onto the RH and then following the same functional localization procedure used in the main analyses (i.e., selecting the top 10% most responsive voxels to the *sentences* > *nonwords* contrast during the localizer task). This approach allows asymmetric patterns of activation across hemispheres at the individual level while continuing to ensure functionally comparable ROIs both within individuals (between hemispheres) and between individuals.

In direct between-hemispheres comparisons, we find significantly reduced length effects in all three experiments in the network overall and in all six regions in the RH relative to the LH, except in the AngG language regions in Experiment 3. We further find significantly reduced length effects for the Jabberwocky conditions of Experiment 2 in the network overall and in all individual regions except the AngG language regions. The length-driven patterns of activation are thus greatly attenuated in RH relative to LH (see Tables S3 and S4 for full testing results). Nonetheless, as shown in Figure A7, the RH homotopic language areas also tend to show length effects in real-

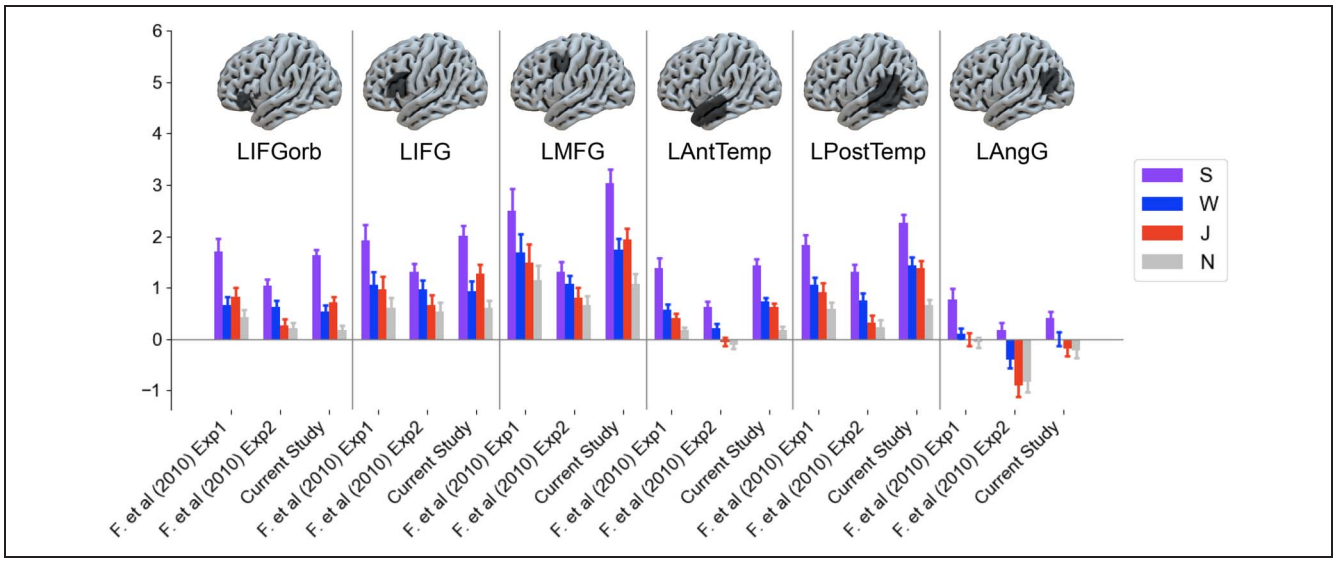

Figure A6. Effect estimates from the sentence (S), word list (W), Jabberwocky sentence (J), and nonword list (N) conditions from Experiments 1 and 2 of Fedorenko and colleagues (2010; left and center) versus the equivalent conditions (S = c12, W = c01, J = jab-c12, N = jab-c01) from the current Experiment 2. Error bars show standard error of the mean across participants.

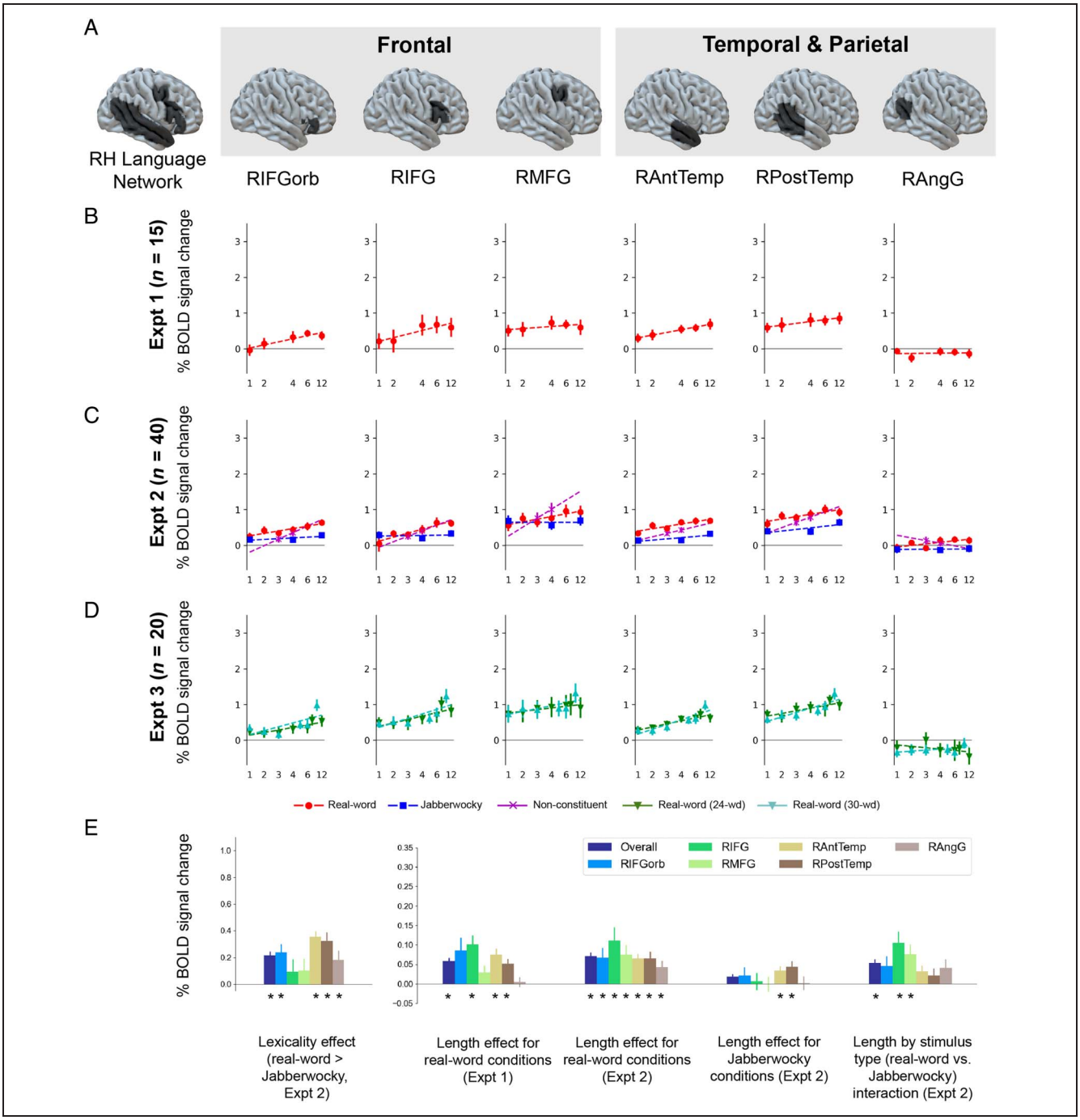

Figure A7. Main results (parallel of Figure 3 of the main article) in the right hemisphere (RH homotopes of LH language areas). (A) Group masks bounding the six right-hemisphere language regions. The top 10% of language-selective voxels are selected within each mask in each participant. (B) Estimated response to each condition of the real-word conditions in Experiment 1 (which did not include Jabberwocky conditions). Responses in three regions (RIFG, RAntTemp, and RPostTemp) increase significantly with constituent length, albeit more weakly than in their left-hemisphere homotopes (Figure 3). (C) Estimated response to each condition of the real-word conditions (replicating Experiment 1), the Jabberwocky conditions, and the nonconstituents conditions in Experiment 2. Responses in all regions increase with constituent length in the real-word conditions and with constituent length in the nonconstituent conditions in all regions but RAngG, but only in temporal regions do responses increase with constituent length in the Jabberwocky conditions. (D) Estimated response to each condition of both the 24-word and 30-word items of Experiment 3, both of which consisted of contiguous real-word chunks that generally did not form syntactic constituents. Responses in all regions but RAngG increase as a function of constituent length to a similar degree to the real-word conditions of Experiments 1 and 2. (E) Key contrasts by fROI (left-to-right): overall lexicality effect (increase in response for real-word over Jabberwocky conditions in Experiment 2, averaging over length); constituent-length effect for real-word conditions in Experiment 1 (slope of the line by participant from B); constituent-length effect for real-word conditions in Experiment 2 (slope of the red line by participant from C); constituent-length effect for Jabberwocky conditions in Experiment 2 (slope of the blue line by participant from C); increase in constituent-length effect in real-word conditions over Jabberwocky in Experiment 2 (difference between the slopes of the red and blue lines by participant from C). Starred bars indicate statistically significant effects by likelihood ratio test (corrected for false discovery rate across fROIs; Benjamini & Yekutieli, 2001). Error bars show standard error of the mean over participants.

Table A4. Significance Tests of Key Contrasts in the Right-hemisphere Language Homotopes with Estimates, Standard Errors, and *t* Values (Respectively Columns β, σ(β), and *t*)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
Constituent length for real-word conditions	1	**Overall**	0.06	0.02	3.53	.006**
	1	RIFGorb	0.09	0.03	2.69	.065
	1	**RIFG**	0.10	0.02	4.23	.006**
	1	RMFG	0.03	0.02	1.62	.375
	1	**RAntTemp**	0.08	0.02	4.96	.003**
	1	**RPostTemp**	0.05	0.01	3.94	.007**
	1	RAngG	0.00	0.01	0.39	1.000
Constituent length for real-word conditions	2	**Overall**	0.07	0.02	3.89	< .001***
	2	**RIFGorb**	0.07	0.03	2.66	.028*
	2	**RIFG**	0.11	0.03	3.23	.012*
	2	**RMFG**	0.08	0.02	3.01	.017*
	2	**RAntTemp**	0.07	0.01	5.74	< .001***
	2	**RPostTemp**	0.07	0.02	3.79	.004**
	2	**RAngG**	0.04	0.02	2.66	.028*
Constituent length for Jabberwocky conditions	2	Overall	0.02	0.01	1.26	.214
	2	RIFGorb	0.02	0.02	0.98	1.000
	2	RIFG	0.01	0.02	0.29	1.000
	2	RMFG	0.00	0.02	−0.05	1.000
	2	**RAntTemp**	0.03	0.01	2.89	.046*
	2	**RPostTemp**	0.04	0.01	3.16	.045*
	2	RAngG	0.00	0.02	0.13	1.000
Lexicality effect (real-word > Jabberwocky)	2	**Overall**	0.22	0.07	3.29	.004**
	2	**RIFGorb**	0.24	0.06	3.80	.002**
	2	RIFG	0.09	0.09	1.02	.773
	2	RMFG	0.10	0.09	1.17	.737
	2	**RAntTemp**	0.35	0.04	7.88	< .001***
	2	**RPostTemp**	0.32	0.06	5.00	< .001***
	2	**RAngG**	0.18	0.07	2.59	.049*
Constituent-Length × Stimulus Type (real-word vs. Jabberwocky) interaction	2	**Overall**	0.05	0.02	2.78	.011*
	2	RIFGorb	0.05	0.03	1.78	.242
	2	**RIFG**	0.11	0.03	3.50	.017*
	2	**RMFG**	0.08	0.03	2.97	.038*
	2	RAntTemp	0.03	0.02	2.07	.219
	2	RPostTemp	0.02	0.02	1.09	.694
	2	RAngG	0.04	0.02	1.83	.242

Table A4. (*continued*)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
Length effect in Experiment 3 (mostly nonconstituents)	3	**Overall**	0.08	0.02	3.43	.005**
	3	**RIFGorb**	0.09	0.02	3.69	.006**
	3	**RIFG**	0.11	0.02	4.62	< .001***
	3	RMFG	0.06	0.03	2.14	.133
	3	**RAntTemp**	0.11	0.02	6.81	< .001***
	3	**RPostTemp**	0.10	0.02	5.85	< .001***
	3	RAngG	−0.01	0.02	−0.33	1.000

p Values are generated by likelihood ratio tests of linear mixed-effects models, with by-fROI results corrected for false discovery rate (FDR) applied over all six fROIs using the Benjamini–Yekutieli procedure (Benjamini & Yekutieli, 2001) using a nominal significance level of $\alpha = .05$. Starred *p* values indicate statistical significance under FDR correction (*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$). Significant regions are shown in **bold** in the fROI column.

word conditions, albeit substantially weaker than those found in the LH (Figure 3 of the main article), with fewer of these effects reaching significance (e.g., the real-word length effect is weak and not significant in the right middle frontal gyrus (RMFG) homotopic language area, whereas it is strong and significant in LMFG). The length effect for Jabberwocky is significant only in the RH temporal language areas, indicating a generally reduced engagement of RH areas in the processing of syntactically well-formed but meaningless stimuli, relative to their LH homotopes. The RH language homotopes thus seem to show similar but attenuated patterns of response to parametric variation of the length of linguistic context.

Table A5. Significance Tests of Laterality Difference (LH − RH) of Key Contrasts with Estimates, Standard Errors, and *t* Values (Respectively Columns β, σ(β), and *t*)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
Laterality difference of constituent length for real-word conditions	1	**Overall**	0.14	0.03	4.32	< .001***
	1	**LIFGorb**	0.20	0.05	3.71	.011*
	1	**LIFG**	0.16	0.04	3.81	.011*
	1	**LMFG**	0.16	0.05	3.42	.015*
	1	**LAntTemp**	0.07	0.02	3.19	.019*
	1	**LPostTemp**	0.15	0.03	5.75	< .001***
	1	**LAngG**	0.08	0.03	3.08	.020*
Laterality difference of constituent length for real-word conditions	2	**Overall**	0.11	0.02	4.96	< .001***
	2	**LIFGorb**	0.17	0.03	6.50	< .001***
	2	**LIFG**	0.12	0.03	3.74	.002**
	2	**LMFG**	0.17	0.02	7.00	< .001***
	2	**LAntTemp**	0.07	0.02	4.67	< .001***
	2	**LPostTemp**	0.09	0.01	6.30	< .001***
	2	**LAngG**	0.05	0.02	2.81	.019*

Table A5. (*continued*)

Contrast	*Experiment*	*fROI*	β	σ(β)	*t*	*p*
Laterality difference of constituent length for Jabberwocky conditions	2	**Overall**	0.09	0.03	3.50	.007**
	2	**LIFGorb**	0.08	0.02	3.99	< .001***
	2	**LIFG**	0.13	0.02	5.50	< .001***
	2	**LMFG**	0.18	0.03	6.61	< .001***
	2	**LAntTemp**	0.05	0.01	3.98	< .001***
	2	**LPostTemp**	0.09	0.01	6.38	< .001***
	2	LAngG	0.00	0.02	0.05	1.000
Laterality difference of lexicality effect (real-word > Jabberwocky)	2	**Overall**	0.53	0.08	6.46	< .001***
	2	**LIFGorb**	0.44	0.09	5.05	< .001***
	2	**LIFG**	0.59	0.11	5.22	< .001***
	2	**LMFG**	0.80	0.09	9.01	< .001***
	2	**LAntTemp**	0.42	0.06	6.67	< .001***
	2	**LPostTemp**	0.62	0.08	8.20	< .001***
	2	**LAngG**	0.31	0.07	4.19	< .001***
Laterality difference of Constituent-Length × Stimulus Type (real-word vs. Jabberwocky) interaction	2	Overall	0.02	0.02	1.42	.173
	2	LIFGorb	0.08	0.03	2.97	.075
	2	LIFG	0.00	0.03	−0.10	1.000
	2	LMFG	−0.01	0.03	−0.23	1.000
	2	LAntTemp	0.02	0.02	1.20	1.000
	2	LPostTemp	0.00	0.02	−0.05	1.000
	2	LAngG	0.05	0.03	1.81	.569
Laterality difference of length effect in Experiment 3 (mostly nonconstituents)	3	**Overall**	0.12	0.02	5.59	< .001***
	3	**LIFGorb**	0.13	0.03	5.07	< .001***
	3	**LIFG**	0.15	0.03	5.04	< .001***
	3	**LMFG**	0.12	0.03	4.46	< .001***
	3	**LAntTemp**	0.10	0.02	5.35	< .001***
	3	**LPostTemp**	0.13	0.02	5.92	< .001***
	3	LAngG	0.07	0.03	2.52	.051

p Values are generated by likelihood ratio tests of linear mixed-effects models, with by-fROI results corrected for false discovery rate (FDR) applied over all six fROIs using the Benjamini–Yekutieli procedure (Benjamini & Yekutieli, 2001) using a nominal significance level of $\alpha = .05$. Starred *p* values indicate statistical significance under FDR correction (*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$). Significant regions are shown in **bold** in the fROI column.

Acknowledgments

We acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research, MIT. For technical support during scanning, we thank Steve Shannon and Atsushi Takahashi. We thank Tamar Regev for help with the ROI projection figures, the audience at the Neurobiology of Language conference in 2020 for helpful discussions of this work, and Rebecca Saxe, Ted Gibson, Nancy Kanwisher, Christophe Pallier, and Stan Dehaene for comments on the article. We also thank Christophe Pallier for sharing the ROI files.

Corresponding author: Cory Shain, Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, or via e-mail: cshain@mit.edu.

Data Availability Statement

All data needed to evaluate the conclusions in the article are publicly available on OSF: https://osf.io/fduve/.

Author Contributions

Cory Shain: Conceptualization; Formal analysis; Investigation; Methodology; Software; Visualization; Writing—Original; Writing—Review & editing. Hope Kean: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Writing—Review & editing. Colton Casto: Data curation; Formal analysis; Investigation; Methodology; Writing—Review & editing. Benjamin Lipkin: Investigation; Methodology. Josef Affourtit: Investigation; Methodology. Matthew Siegelman: Investigation; Methodology. Francis Mollica: Conceptualization; Methodology; Project administration; Supervision; Writing—Review & editing. Evelina Fedorenko: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Supervision; Writing—Original draft; Writing—Review & editing.

Funding Information

E. F. was supported by National Institutes of Health (https://dx.doi.org/10.13039/100000002), grant numbers: R01-DC016607, R01-DC016950, and U01-NS121471, and by the funds from the McGovern Institute for Brain Research, Brain and Cognitive Sciences Department (https://dx.doi.org/10.13039/100019335), and the Simons Center for the Social Brain (https://dx.doi.org/10.13039/100018792).

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

REFERENCES

Aliko, S., Wang, B., Small, S. L., & Skipper, J. I. (2023). The entire brain, more or less is at work: 'Language regions' are artefacts of averaging. *bioRxiv*, 555886. https://doi.org/10.1101/2023.09.01.555886

Amit, E., Hoeflin, C., Hamzah, N., & Fedorenko, E. (2017). An asymmetrical relationship between verbal and visual thinking: Converging evidence from behavior and fMRI. *Neuroimage*, *152*, 619–627. https://doi.org/10.1016/j.neuroimage.2017.03.029, PubMed: 28323162

Anderson, A. J., Kiela, D., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., et al. (2021). Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience*, *41*, 4100–4119. https://doi.org/10.1523/JNEUROSCI.1152-20.2021, PubMed: 33753548

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage*, *26*, 839–851. https://doi.org/10.1016/j.neuroimage.2005.02.018, PubMed: 15955494

Baggio, G. (2021). Compositionality in a parallel architecture for language processing. *Cognitive Science*, *45*, e12949. https://doi.org/10.1111/cogs.12949, PubMed: 34018238

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, *26*, 1338–1367. https://doi.org/10.1080/01690965.2010.542671

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate. *Psychological Review*, *113*, 628–647. https://doi.org/10.1037/0033-295X.113.3.628, PubMed: 16802884

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 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*, 567–574. https://doi.org/10.1080/23273798.2015.1123281, PubMed: 27525290

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences, U.S.A.*, *108*, 4429–4434. https://doi.org/10.1073/pnas.1014818108, PubMed: 21368161

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*, 1165–1188. https://doi.org/10.1214/aos/1013699998

Berwick, R. C., Beckers, G. J. L., Okanoya, K., & Bolhuis, J. J. (2012). A bird's eye view of human language evolution. *Frontiers in Evolutionary Neuroscience*, *4*, 5. https://doi.org/10.3389/fnevo.2012.00005, PubMed: 22518103

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*, 2767–2796. https://doi.org/10.1093/cercor/bhp055, PubMed: 19329570

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, PubMed: 26666896

Blank, I. A., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows.

Neuroimage, *219*, 116925. https://doi.org/10.1016/j.neuroimage.2020.116925, PubMed: 32407994
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*, 1105–1118. https://doi.org/10.1152/jn.00884.2013, PubMed: 24872535
Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences, U.S.A.*, *113*, 10818–10823. https://doi.org/10.1073/pnas.1605782113, PubMed: 27621455
Bolhuis, J. J., Tattersall, I., Chomsky, N., & Berwick, R. C. (2014). How could language have evolved? *PLoS Biology*, *12*, e1001934. https://doi.org/10.1371/journal.pbio.1001934, PubMed: 25157536
Bonner, M. F., Peelle, J. E., Cook, P. A., & Grossman, M. (2013). Heteromodal conceptual processing in the angular gyrus. *Neuroimage*, *71*, 175–186. https://doi.org/10.1016/j.neuroimage.2013.01.006, PubMed: 23333416
Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013). Reconciling time, space and function: A new dorsal-ventral stream model of sentence comprehension. *Brain and Language*, *125*, 60–76. https://doi.org/10.1016/j.bandl.2013.01.010, PubMed: 23454075
Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S. L., & Rauschecker, J. P. (2015). Neurobiological roots of language in primate audition: Common computational properties. *Trends in Cognitive Sciences*, *19*, 142–150. https://doi.org/10.1016/j.tics.2014.12.008, PubMed: 25600585
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*, 1415–1448. https://doi.org/10.1152/jn.00753.2019, PubMed: 32965153
Branco, P., Seixas, D., & Castro, S. L. (2020). Mapping language with resting-state functional magnetic resonance imaging: A study on the functional profile of the language network. *Human Brain Mapping*, *41*, 545–560. https://doi.org/10.1002/hbm.24821, PubMed: 31609045
Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLoS One*, *14*, e0207741. https://doi.org/10.1371/journal.pone.0207741, PubMed: 30650078
Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*, 163–173. https://doi.org/10.1016/j.bandl.2010.04.002, PubMed: 20472279
Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157–158*, 81–94. https://doi.org/10.1016/j.bandl.2016.04.008, PubMed: 27208858
Cappa, S. F. (2012). Imaging semantics and syntax. *Neuroimage*, *61*, 427–431. https://doi.org/10.1016/j.neuroimage.2011.10.006, PubMed: 22019859
Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language*, *3*, 572–582. https://doi.org/10.1016/0093-934x(76)90048-1, PubMed: 974731
Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Disentangling syntax and semantics in the brain with deep networks. *International Conference on Machine Learning*, *139*, 1336–1348.
Chen, X., Affourtit, J., Ryskin, R., Regev, T. I., Norman-Haignere, S., Jouravlev, O., et al. (2023). The human language system, including its inferior frontal component in "Broca's area," does not support music perception. *Cerebral Cortex*, *33*, 7904–7929. https://doi.org/10.1093/cercor/bhad087, PubMed: 37005063
Chen, E., Gibson, E., & Wolf, F. (2005). Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, *52*, 144–169. https://doi.org/10.1016/j.jml.2004.10.001
Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., & Cox, R. W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. *Neuroimage*, *73*, 176–190. https://doi.org/10.1016/j.neuroimage.2013.01.047, PubMed: 23376789
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, N. (1995a). Bare phrase structure. *Government and Binding Theory and the Minimalist Program*, 383–439.
Chomsky, N. (1995b). *The minimalist program*. Cambridge, MA: MIT Press.
Church, A. (1932). A set of postulates for the foundation of logic. *Annals of Mathematics*, *33*, 346–366. https://doi.org/10.2307/1968337
Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage*, *9*, 179–194. https://doi.org/10.1006/nimg.1998.0395, PubMed: 9931268
Davey, J., Thompson, H. E., Hallam, G., Karapanagiotidis, T., Murphy, C., De Caso, I., et al. (2016). Exploring the role of the posterior middle temporal gyrus in semantic cognition: Integration of anterior temporal lobe with executive processes. *Neuroimage*, *137*, 165–177. https://doi.org/10.1016/j.neuroimage.2016.05.051, PubMed: 27236083
Davis, C. P., & Yee, E. (2019). Features, labels, space, and time: Factors supporting taxonomic relationships in the anterior temporal lobe and thematic relationships in the angular gyrus. *Language, Cognition and Neuroscience*, *34*, 1347–1357. https://doi.org/10.1080/23273798.2018.1479530
Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, *25*, 4596–4609. https://doi.org/10.1093/cercor/bhv111, PubMed: 26048954
Dehaene, S. (2019). Human singularity and symbolic tree structures: The demodularization hypothesis. In W. Singer, T. J. Sejnowski, & P. Rakic (Eds.), *The neocortex* (Vol. 27, p. 293). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/12593.003.0021
Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: A hypothesis about human singularity. *Trends in Cognitive Sciences*, *26*, 751–766. https://doi.org/10.1016/j.tics.2022.06.010, PubMed: 35933289
Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees. *Neuron*, *88*, 2–19. https://doi.org/10.1016/j.neuron.2015.09.019, PubMed: 26447569
Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210. https://doi.org/10.1016/j.cognition.2008.07.008, PubMed: 18930455
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*, 4536–4550. https://doi.org/10.1523/JNEUROSCI.2036-19.2020, PubMed: 32317387
Duffau, H., Moritz-Gasser, S., & Mandonnet, E. (2014). A re-examination of neural basis of language processing:

Proposal of a dynamic hodotopical model from data provided by brain stimulation mapping during picture naming. *Brain and Language*, *131*, 1–10. https://doi.org/10.1016/j.bandl.2013.05.011, PubMed: 23866901

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 199–209). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1024

Fedorenko, E. (2021). The early origins and the growing popularity of the individual-subject analytic approach in human neuroscience. *Current Opinion in Behavioral Sciences*, *40*, 105–112. https://doi.org/10.1016/j.cobeha.2021.02.023

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, U.S.A.*, *108*, 16428–16433. https://doi.org/10.1073/pnas.1112937108, PubMed: 21885736

Fedorenko, E., & Blank, I. A. (2020). Broca's area is not a natural kind. *Trends in Cognitive Sciences*, *24*, 270–284. https://doi.org/10.1016/j.tics.2020.01.001, PubMed: 321605645

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, PubMed: 32569894

Fedorenko, E., Duncan, J., & Kanwisher, N. (2012). Language-selective and domain-general regions lie side by side within Broca's area. *Current Biology*, *22*, 2059–2062. https://doi.org/10.1016/j.cub.2012.09.011, PubMed: 23063434

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*, 1177–1194. https://doi.org/10.1152/jn.00032.2010, PubMed: 20410363

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, *25*, 289–312. https://doi.org/10.1038/s41583-024-00802-4, PubMed: 38609551

Fedorenko, E., & Kanwisher, N. (2009). Neuroimaging of language: Why hasn't a clearer picture emerged? *Language and Linguistics Compass*, *3*, 839–865. https://doi.org/10.1111/j.1749-818X.2009.00143.x

Fedorenko, E., Nieto-Castañon, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*, 499–513. https://doi.org/10.1016/j.neuropsychologia.2011.09.014, PubMed: 21945850

Fedorenko, E., Nieto-Castañón, A., & Kanwisher, N. (2012). Syntactic processing in the human brain: What we know, what we don't know, and a suggestion for how to proceed. *Brain and Language*, *120*, 187–207. https://doi.org/10.1016/j.bandl.2011.01.001, PubMed: 21334056

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., et al. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences, U.S.A.*, *113*, E6256–E6262. https://doi.org/10.1073/pnas.1612132113, PubMed: 27671642

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, *18*, 120–126. https://doi.org/10.1016/j.tics.2013.12.006, PubMed: 24440115

Fedorenko, E., & Varley, R. (2016). Language and thought are not the same thing: Evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, *1369*, 132–153. https://doi.org/10.1111/nyas.13046, PubMed: 27096882

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., et al. (2016). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex*, *26*, 2018–2034. https://doi.org/10.1093/cercor/bhv020, PubMed: 25750259

Fitch, W. T. (2014). Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. *Physics of Life Reviews*, *11*, 329–364. https://doi.org/10.1016/j.plrev.2014.04.005, PubMed: 24969660

Fitch, W. T., & Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, *1316*, 87–104. https://doi.org/10.1111/nyas.12406, PubMed: 24697242

Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/4737.001.0001

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In D. Reitter & R. Levy (Eds.), *Proceedings of the 3rd workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)* (pp. 61–69). Montréal, Canada: Association for Computational Linguistics.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*, 829–834. https://doi.org/10.1177/0956797611409589, PubMed: 21586764

Frankland, S. M., & Greene, J. D. (2020). Concepts and compositionality: In search of the brain's language of thought. *Annual Review of Psychology*, *71*, 273–303. https://doi.org/10.1146/annurev-psych-122216-011829, PubMed: 31550985

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and performance 12: The psychology of reading* (pp. 559–586). Erlbaum.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, *91*, 1357–1392. https://doi.org/10.1152/physrev.00006.2011, PubMed: 22013214

Friederici, A. D. (2017). *Language in our brain: The origins of a uniquely human capacity*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/11173.001.0001

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

Friston, K., & Buzsáki, G. (2016). The functional anatomy of time: What and when in the brain. *Trends in Cognitive Sciences*, *20*, 500–511. https://doi.org/10.1016/j.tics.2016.05.001, PubMed: 27261057

Frost, M. A., & Goebel, R. (2012). Measuring structural-functional correspondence: Spatial variability of specialised brain regions after macro-anatomical alignment. *Neuroimage*, *59*, 1369–1381. https://doi.org/10.1016/j.neuroimage.2011.08.035, PubMed: 21875671

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., et al. (2021). The natural stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*, 63–77. https://doi.org/10.1007/s10579-020-09503-7, PubMed: 34720781

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. P. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95–126). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/3654.003.0008

Giglio, L., Ostarek, M., Weber, K., & Hagoort, P. (2022). Commonalities and asymmetries in the neurobiological infrastructure for language production and comprehension. *Cerebral Cortex*, *32*, 1405–1418. https://doi.org/10.1093/cercor/bhab287, PubMed: 34491301

Gilmore, R. O., Diaz, M. T., Wyble, B. A., & Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1396*, 5–18. https://doi.org/10.1111/nyas.13325, PubMed: 28464561

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*, 171–178. https://doi.org/10.1038/nature18933, PubMed: 27437579

Goldberg, A. (2005). *Constructions at work: The nature of generalization in language*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199268511.001.0001

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In A. Sayeed, C. Jacobs, T. Linzen, & M. van Schijndel (Eds.), *Proceedings of the 8th workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)* (pp. 10–18). Salt Lake City, Utah: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0102

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1411–1423. https://doi.org/10.1037/0278-7393.27.6.1411, PubMed: 11713876

Goucha, T., & Friederici, A. D. (2015). The language skeleton after dissecting meaning: A functional segregation within Broca's area. *Neuroimage*, *114*, 294–302. https://doi.org/10.1016/j.neuroimage.2015.04.011, PubMed: 25871627

Graff, D., Kong, J., Chen, K., & Maeda, K. (2007). *English gigaword third edition LDC2007T07*. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2007T07

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences, U.S.A.*, *100*, 253–258. https://doi.org/10.1073/pnas.0135058100, PubMed: 12506194

Grodner, D. J., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*, 261–290. https://doi.org/10.1207/s15516709cog0000_7, PubMed: 21702774

Grodzinsky, Y., Pieperhoff, P., & Thompson, C. (2021). Stable brain loci for the processing of complex syntax: A review of the current neuroimaging evidence. *Cortex*, *142*, 252–271. https://doi.org/10.1016/j.cortex.2021.06.003, PubMed: 34303116

Hage, S. R., & Nieder, A. (2016). Dual neural network model for the evolution of speech and language. *Trends in Neurosciences*, *39*, 813–829. https://doi.org/10.1016/j.tins.2016.10.006, PubMed: 27884462

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, *9*, 416–423. https://doi.org/10.1016/j.tics.2005.07.004, PubMed: 16054419

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 159–166.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*, 643–672. https://doi.org/10.1207/s15516709cog0000_64, PubMed: 21702829

Hariri, A. R. (2009). The neurobiology of individual differences in complex behavioral traits. *Annual Review of Neuroscience*, *32*, 225–247. https://doi.org/10.1146/annurev.neuro.051508.135335, PubMed: 19400720

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, *28*, 2539–2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008, PubMed: 18322098

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics* (pp. 690–696).

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences, U.S.A.*, *119*, e2201968119. https://doi.org/10.1073/pnas.2201968119, PubMed: 35921434

Henderson, J. M., Choi, W., Luke, S. G., & Desai, R. H. (2015). Neural correlates of fixation duration in natural reading: Evidence from fixation-related fMRI. *Neuroimage*, *119*, 390–397. https://doi.org/10.1016/j.neuroimage.2015.06.072, PubMed: 26151101

Hertrich, I., Dietrich, S., & Ackermann, H. (2016). The role of the supplementary motor area for speech and language processing. *Neuroscience & Biobehavioral Reviews*, *68*, 602–610. https://doi.org/10.1016/j.neubiorev.2016.06.030, PubMed: 27343998

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*, 393–402. https://doi.org/10.1038/nrn2113, PubMed: 17431404

Holmes, A. P., & Friston, K. J. (1998). Generalisability, random effects & population inference. *Neuroimage*, *7*, S754. https://doi.org/10.1016/S1053-8119(18)31587-8

Hu, J., Small, H., Kean, H., Takahashi, A., Zekelman, L., Kleinman, D., et al. (2023). Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, *30*, 4384–4404. https://doi.org/10.1093/cercor/bhac350, PubMed: 36130104

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of Cognitive Neuroscience*, *18*, 665–679. https://doi.org/10.1162/jocn.2006.18.4.665

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*, 453–458. https://doi.org/10.1038/nature17637, PubMed: 27121839

Ivanova, A. A., Mineroff, Z., Zimmerer, V., Kanwisher, N., Varley, R., & Fedorenko, E. (2021). The language network is recruited but not required for nonverbal event semantics. *Neurobiology of Language*, *2*, 176–201. https://doi.org/10.1162/nol_a_00030, PubMed: 37216147

Ivanova, A. A., Siegelman, M., Cheung, C., Pongos, A., Kean, H., & Fedorenko, E. (2020). The effect of task on brain activity during sentence processing. *12th Annual Meeting of the Society for the Neurobiology of Language (SNL20)*.

Jackendoff, R. S. (1990). *Semantic structures*. Cambridge, MA: MIT Press.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, *32*, 37–55. https://doi.org/10.1023/a:1021933015362, PubMed: 12647562

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences, U.S.A.*, *107*, 11163–11170. https://doi.org/10.1073/pnas.1005062107, PubMed: 20484679

Kaplan, R. M., & Bresnan, J. (1982). Lexical functional grammar: A formal system for grammatical representation. In J. Bresnan (Ed.), *The mental representation of grammatical relations* (pp. 173–281). Cambridge, MA: MIT Press.

Kauf, C., Tuckute, G., Levy, R., Andreas, J., & Fedorenko, E. (2024). Lexical semantic content, not syntactic structure, is the main contributor to ANN-brain similarity of fMRI responses in the language network. *bioRxiv*. https://doi.org/10.1101/2023.05.05.539646, PubMed: 37205405

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*, 223–237. https://doi.org/10.1093/cercor/11.3.223, PubMed: 11230094

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, *41*, 1–12. https://doi.org/10.3758/BRM.41.1.12, PubMed: 19182118

Kempen, G. (2014). Prolegomena to a neurocomputational architecture for human grammatical encoding and decoding. *Neuroinformatics*, *12*, 111–142. https://doi.org/10.1007/s12021-013-9191-4, PubMed: 23872869

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*, 627–633. https://doi.org/10.3758/BRM.42.3.627, PubMed: 20805584

Koechlin, E., & Jubault, T. (2006). Broca's area and the hierarchical organization of human behavior. *Neuron*, *50*, 963–974. https://doi.org/10.1016/j.neuron.2006.05.017, PubMed: 16772176

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*, 535–540. https://doi.org/10.1038/nn.2303, PubMed: 19396166

Lashley, K. (1951). The problem of serial order in behaviour. *Cerebral Mechanisms in Behaviour*, 112–136.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*, 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011, PubMed: 21414912

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006, PubMed: 17662975

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 375–419. https://doi.org/10.1207/s15516709cog0000_25, PubMed: 21702779

Li, X. (1988). Effects of contextual cues on inferring and remembering meanings of new words. *Applied Linguistics*, *9*, 402–413. https://doi.org/10.1093/applin/9.4.402

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., et al. (2022). Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Scientific Data*, *9*, 529. https://doi.org/10.1038/s41597-022-01645-3, PubMed: 36038572

Lopopolo, A., Frank, S. L., van den Bosch, A., & Willems, R. M. (2017). Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLoS One*, *12*, e0177794. https://doi.org/10.1371/journal.pone.0177794, PubMed: 28542396

Lopopolo, A., van den Bosch, A., Petersson, K.-M., & Willems, R. M. (2021). Distinguishing syntactic operations in the brain: Dependency and phrase-structure parsing. *Neurobiology of Language*, *2*, 152–175. https://doi.org/10.1162/nol_a_00029, PubMed: 37213416

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703. https://doi.org/10.1037/0033-295x.101.4.676, PubMed: 7984711

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, PubMed: 27261158

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304–316. https://doi.org/10.3102/0013189X14545513

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffman, M., Mineroff, Z., et al. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*, 1014–1019. https://doi.org/10.1038/s41593-022-01114-5, PubMed: 35856094

Malik-Moraleda, S., Jouravlev, O., Taliaferro, M., Mineroff, Z., Cucu, T., Mahowald, K., et al. (2024). Functional characterization of the language network of polyglots and hyperpolyglots with precision fMRI. *Cerebral Cortex*, *34*, bhae049. https://doi.org/10.1093/cercor/bhae049, PubMed: 38466812

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*, 313–330.

Martin, K. C., Seydell-Greenwald, A., Berl, M. M., Gaillard, W. D., Turkeltaub, P. E., & Newport, E. L. (2022). A weak shadow of early life language processing persists in the right hemisphere of the mature brain. *Neurobiology of Language*, *3*, 364–385. https://doi.org/10.1162/nol_a_00069, PubMed: 35686116

Matchin, W., Brodbeck, C., Hammerly, C., & Lau, E. (2019). The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG. *Human Brain Mapping*, *40*, 663–678. https://doi.org/10.1002/hbm.24403, PubMed: 30259599

Matchin, W., Hammerly, C., & Lau, E. (2017). The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *Cortex*, *88*, 106–123. https://doi.org/10.1016/j.cortex.2016.12.010, PubMed: 28088041

Matchin, W., & Hickok, G. (2020). The cortical organization of syntax. *Cerebral Cortex*, *30*, 1481–1498. https://doi.org/10.1093/cercor/bhz180, PubMed: 31670779

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L., & Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, *5*, 467–479. https://doi.org/10.1162/jocn.1993.5.4.467, PubMed: 23964919

Merlin, G., & Toneva, M. (2022). Language models and brain alignment: Beyond word-level semantics and prediction. *ArXiv Preprint* ArXiv:2212.00596. https://doi.org/10.48550/arXiv.2212.00596

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv*, *abs/1301.3*, 1–12. https://doi.org/10.48550/arXiv.1301.3781

Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., et al. (2020). Composition is the core driver of the language-selective network. *Neurobiology of Language*, *1*, 104–134. https://doi.org/10.1162/nol_a_00005, PubMed: 36794007

Montague, R. (1970). Universal grammar. *Theoria*, *36*, 373–398. https://doi.org/10.1111/j.1755-2567.1970.tb00434.x

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2012). Thought beyond language: Neural dissociation of algebra and natural language. *Psychological Science*, *23*, 914–922. https://doi.org/10.1177/0956797612437427, PubMed: 22760883

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., et al. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing.

Proceedings of the National Academy of Sciences, U.S.A., *114*, E3669–E3678. https://doi.org/10.1073/pnas.1701590114, PubMed: 28416691
Nguyen, L., van Schijndel, M., & Schuler, W. (2012). Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of COLING 2012* (pp. 2125–2140).
Nieto-Castanon, A. (2020). *Handbook of functional connectivity magnetic resonance imaging methods in CONN*. Hilbert Press. https://doi.org/10.56441/hilbertpress.2207.6598
Nieto-Castañón, A., & Fedorenko, E. (2012). Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *Neuroimage*, *63*, 1646–1669. https://doi.org/10.1016/j.neuroimage.2012.06.065, PubMed: 22784644
Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105–1107. https://doi.org/10.1038/nn.2886, PubMed: 21878926
Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., et al. (2016). *Universal dependencies v1: A multilingual treebank collection*. LREC.
Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective & Behavioral Neuroscience*, *5*, 263–281. https://doi.org/10.3758/cabn.5.3.263, PubMed: 16396089
Oh, B.-D., & Schuler, W. (2021). Contributions of propositional content and syntactic category information in sentence processing. *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 241–250). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.cmcl-1.28
Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. *Neuropsychologia*, *9*, 97–113. https://doi.org/10.1016/0028-3932(71)90067-4, PubMed: 5146491
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. https://doi.org/10.1126/science.aac4716, PubMed: 26315443
Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*, 903–911. https://doi.org/10.1038/nn.4021, PubMed: 25984889
Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences, U.S.A.*, *108*, 2522–2527. https://doi.org/10.1073/pnas.1018711108, PubMed: 21224415
Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. https://doi.org/10.1177/1745691612465253, PubMed: 26168108
Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*, 674–681. https://doi.org/10.1038/nn1082, PubMed: 12830158
Pattamadilok, C., Dehaene, S., & Pallier, C. (2016). A role for left inferior frontal and posterior superior temporal cortex in extracting a syntactic tree from a sentence. *Cortex*, *75*, 44–55. https://doi.org/10.1016/j.cortex.2015.11.012, PubMed: 26709465
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*, 1244–1265. https://doi.org/10.1152/jn.00619.2018, PubMed: 30601693
Petkov, C. I., & Jarvis, E. D. (2012). Birds, primates, and spoken language origins: Behavioral phenotypes and neurobiological substrates. *Frontiers in Evolutionary Neuroscience*, *4*, 12. https://doi.org/10.3389/fnevo.2012.00012, PubMed: 22912615
Philippi, C. L., Tranel, D., Duff, M., & Rudrauf, D. (2015). Damage to the default mode network disrupts autobiographical memory retrieval. *Social Cognitive and Affective Neuroscience*, *10*, 318–326. https://doi.org/10.1093/scan/nsu070, PubMed: 24795444
Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.
Price, A. R., Bonner, M. F., Peelle, J. E., & Grossman, M. (2015). Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *Journal of Neuroscience*, *35*, 3276–3284. https://doi.org/10.1523/JNEUROSCI.3446-14.2015, PubMed: 25698762
Price, A. R., Peelle, J. E., Bonner, M. F., Grossman, M., & Hamilton, R. H. (2016). Causal evidence for a mechanism of semantic integration in the angular gyrus as revealed by high-definition transcranial direct current stimulation. *Journal of Neuroscience*, *36*, 3829–3838. https://doi.org/10.1523/JNEUROSCI.3120-15.2016, PubMed: 27030767
Pylkkänen, L., & McElree, B. (2006). The syntax-semantics interface: On-line composition of sentence meaning. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 539–579). Elsevier. https://doi.org/10.1016/B978-012369374-7/50015-8
Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences, U.S.A.*, *98*, 676–682. https://doi.org/10.1073/pnas.98.2.676, PubMed: 11209064
Rasmussen, N. E., & Schuler, W. (2018). Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive Science*, *42 Suppl 4*, 1009–1042. https://doi.org/10.1111/cogs.12511, PubMed: 28763111
Reddy, A. J., & Wehbe, L. (2021). Can fMRI reveal the representation of syntactic structure in the brain? In *Advances in neural information processing systems* (Vol. 34, pp. 9843–9856).
Regev, T. I., Affourtit, J., Chen, X., Schipper, A. E., Bergen, L., Mahowald, K., et al. (2024). High-level language brain regions are sensitive to sub-lexical regularities. *Cerebral Cortex*, *34*, bhae077. https://doi.org/10.1093/cercor/bhae077, PubMed: 38494886
Regev, T. I., Casto, C., Hosseini, E. A., Adamek, M., Ritaccio, A. L., Willie, J. T., et al. (2023). Neural populations in the language network differ in the size of their temporal receptive windows. *bioRxiv*, 522216. https://doi.org/10.1101/2022.12.30.522216
Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, *15*, 1261–1269. https://doi.org/10.1093/cercor/bhi009, PubMed: 15635062
Rodd, J. M., Longe, O. A., Randall, B., & Tyler, L. K. (2010). The functional organisation of the fronto-temporal language system: Evidence from syntactic and semantic ambiguity. *Neuropsychologia*, *48*, 1324–1335. https://doi.org/10.1016/j.neuropsychologia.2009.12.035, PubMed: 20038434
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, PubMed: 25576690
Rogalsky, C., & Hickok, G. (2009). Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral Cortex*, *19*, 786–796. https://doi.org/10.1093/cercor/bhn126, PubMed: 18669589

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *Neuroimage*, *30*, 1088–1096. https://doi.org/10.1016/j.neuroimage.2005.12.062, PubMed: 16635578

Schuler, W., & Wheeler, A. (2014). Cognitive compositional semantics using continuation dependencies. In *Proceedings of the third joint conference on lexical and computational semantics (*SEM 2014)* (pp. 141–150). Dublin, Ireland: Association for Computational Linguistics and Dublin City University. https://doi.org/10.3115/v1/S14-1018

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*, 167–176. https://doi.org/10.1080/17588928.2016.1201466, PubMed: 27386919

Seghier, M. L. (2013). The angular gyrus: Multiple functions and multiple subdivisions. *Neuroscientist*, *19*, 43–61. https://doi.org/10.1177/1073858412440596, PubMed: 22547530

Shain, C., Blank, I. A., Fedorenko, E., Gibson, E., & Schuler, W. (2022). Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *Journal of Neuroscience*, *42*, 7412–7430. https://doi.org/10.1523/JNEUROSCI.1894-21.2022, PubMed: 36002263

Shain, C., Blank, I. A., van Schijndel, M., 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, PubMed: 31874149

Shain, C., Paunov, A., Chen, X., Lipkin, B., & Fedorenko, E. (2023). No evidence of theory of mind reasoning in the human language network. *Cerebral Cortex*, *33*, 6299–6319. https://doi.org/10.1093/cercor/bhac505, PubMed: 36585774

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., & Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* (pp. 49–58). Osaka, Japan: The COLING 2016 Organizing Committee.

Shashidhara, S., Spronkers, F. S., & Erez, Y. (2020). Individual-subject functional localization increases univariate activation but not multivariate pattern discriminability in the "multiple-demand" frontoparietal network. *Journal of Cognitive Neuroscience*, *32*, 1348–1368. https://doi.org/10.1162/jocn_a_01554, PubMed: 32108555

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80. https://doi.org/10.1177/1745691613514755, PubMed: 26173243

Skeide, M. A., Brauer, J., & Friederici, A. D. (2016). Brain functional and structural predictors of language performance. *Cerebral Cortex*, *26*, 2127–2139. https://doi.org/10.1093/cercor/bhv042, PubMed: 25770126

Skipper, J. I. (2015). The NOLB model: A model of the natural organization of language and the brain. In R. M. Willems (Ed.), *Cognitive neuroscience of natural language use* (pp. 101–134). Cambridge University Press. https://doi.org/10.1017/CBO9781107323667.006

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319. https://doi.org/10.1016/j.cognition.2013.02.013, PubMed: 23747651

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. *Mapping the mind: Domain specificity in cognition and culture* (pp. 39–67). https://doi.org/10.1017/CBO9780511752902.003

Steedman, M. (2001). *The syntactic process*. Cambridge, MA: MIT Press/Bradford Books. https://doi.org/10.7551/mitpress/6591.001.0001

Tahmasebi, A. M., Davis, M. H., Wild, C. J., Rodd, J. M., Hakyemez, H., Abolmaesumi, P., et al. (2012). Is the link between anatomical structure and function equally strong at all cognitive levels of processing? *Cerebral Cortex*, *22*, 1593–1603. https://doi.org/10.1093/cercor/bhr205, PubMed: 21893681

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, *26*, 858–866. https://doi.org/10.1038/s41593-023-01304-9, PubMed: 37127759

Tie, Y., Rigolo, L., Norton, I. H., Huang, R. Y., Wu, W., Orringer, D., et al. (2014). Defining language networks from resting-state fMRI for surgical planning—A feasibility study. *Human Brain Mapping*, *35*, 1018–1030. https://doi.org/10.1002/hbm.22231, PubMed: 23288627

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., et al. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behavior*, *8*, 544–561. https://doi.org/10.1038/s41562-023-01783-7, PubMed: 38172630

van Schijndel, M., Exley, A., & Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, *5*, 522–540. https://doi.org/10.1111/tops.12034, PubMed: 23765642

van Schijndel, M., & Schuler, W. (2015). Hierarchic syntax improves reading time prediction. *Proceedings of NAACL-HLT 2015* (pp. 1597–1605). Denver, Colorado: Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1183

Vandenberghe, R., Nobre, A. C., & Price, C. J. (2002). The response of left temporal cortex to sentences. *Journal of Cognitive Neuroscience*, *14*, 550–560. https://doi.org/10.1162/08989290260045800, PubMed: 12126497

Vázquez-Rodríguez, B., Suárez, L. E., Markello, R. D., Shafiei, G., Paquola, C., Hagmann, P., et al. (2019). Gradients of structure-function tethering across neocortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *116*, 21219–21227. https://doi.org/10.1073/pnas.1903403116, PubMed: 31570622

Vincent, J. L., Snyder, A. Z., Fox, M. D., Shannon, B. J., Andrews, J. R., Raichle, M. E., et al. (2006). Coherent spontaneous activity identifies a hippocampal-parietal memory network. *Journal of Neurophysiology*, *96*, 3517–3531. https://doi.org/10.1152/jn.00048.2006, PubMed: 16899645

Wang, L., Uhrig, L., Jarraya, B., & Dehaene, S. (2015). Representation of numerical and sequential patterns in macaque and human brains. *Current Biology*, *25*, 1966–1974. https://doi.org/10.1016/j.cub.2015.06.035, PubMed: 26212883

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*, 2506–2516. https://doi.org/10.1093/cercor/bhv075, PubMed: 25903464

Wilson, S. M., DeMarco, A. T., Henry, M. L., Gesierich, B., Babiak, M., Mandelli, M. L., et al. (2014). What role does the anterior temporal lobe play in sentence-level processing? Neural correlates of syntactic processing in semantic variant primary progressive aphasia. *Journal of Cognitive Neuroscience*, *26*, 970–985. https://doi.org/10.1162/jocn_a_00550, PubMed: 24345172

Wilson, S. M., Entrup, J. L., Schneck, S. M., Onuscheck, C. F., Levy, D. F., Rahman, M., et al. (2023). Recovery from aphasia in the first year after stroke. *Brain*, *146*, 1021–1039. https://doi.org/10.1093/brain/awac129, PubMed: 35388420

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*, 238–252. https://doi.org/10.1162/089892904322984535, PubMed: 15068594

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122. https://doi.org/10.1177/1745691617693393, PubMed: 28841086

Zaccarella, E., & Friederici, A. D. (2015). Merge in the human brain: A sub-region based functional investigation in the left pars opercularis. *Frontiers in Psychology*, *6*, 1818. https://doi.org/10.3389/fpsyg.2015.01818, PubMed: 26640453

Zaccarella, E., Schell, M., & Friederici, A. D. (2017). Reviewing the functional basis of the syntactic merge mechanism for language: A coordinate-based activation likelihood estimation meta-analysis. *Neuroscience & Biobehavioral Reviews*, *80*, 646–656. https://doi.org/10.1016/j.neubiorev.2017.06.011, PubMed: 28743620