Cerebral Cortex, 2021;00: 1–18

doi: 10.1093/cercor/bhab065
Original Article

ORIGINAL ARTICLE

Incremental Language Comprehension Difficulty Predicts Activity in the Language Network but Not the Multiple Demand Network

Leila Wehbe[1], Idan Asher Blank[2,3], Cory Shain[4], Richard Futrell[2,5], Roger Levy[2,6], Titus Malsburg von der[2,7], Nathaniel Smith[6], Edward Gibson[2] and Evelina Fedorenko[2,8]

[1]Carnegie Mellon University, Machine Learning Department PA 15213, USA, [2]Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences MA 02139, USA, [3]University of California Los Angeles, Department of Psychology CA 90095, USA, [4]Ohio State University, Department of Linguistics OH 43210, USA, [5]University of California Irvine, Department of Linguistics CA 92697, USA, [6]University of California San Diego, Department of Linguistics CA 92161, USA, [7]University of Stuttgart, Institute of Linguistics, 70049 Stuttgart, Germany and [8]Massachusetts Institute of Technology, McGovern Institute for Brain ResearchMA 02139, USA

Address correspondence to email: Leila Wehbe. Email: lwehbe@cmu.edu; or Ev Fedorenko. Email: evelina9@mit.edu.

Abstract

What role do domain-general executive functions play in human language comprehension? To address this question, we examine the relationship between behavioral measures of comprehension and neural activity in the domain-general "multiple demand" (MD) network, which has been linked to constructs like attention, working memory, inhibitory control, and selection, and implicated in diverse goal-directed behaviors. Specifically, functional magnetic resonance imaging data collected during naturalistic story listening are compared with theory-neutral measures of online comprehension difficulty and incremental processing load (reading times and eye-fixation durations). Critically, to ensure that variance in these measures is driven by features of the linguistic stimulus rather than reflecting participant- or trial-level variability, the neuroimaging and behavioral datasets were collected in nonoverlapping samples. We find no behavioral-neural link in functionally localized MD regions; instead, this link is found in the domain-specific, fronto-temporal "core language network," in both left-hemispheric areas and their right hemispheric homotopic areas. These results argue against strong involvement of domain-general executive circuits in language comprehension.

Key words: eye-tracking, fMRI, neural activity, psycholinguistic theories, self-paced reading

Introduction

Human language comprehension encompasses a host of complex computations, from perceptual (auditory, visual or, in the case of Braille, haptic) processing, to word recognition, to recovering the semantic and syntactic dependency structures linking words together, to constructing discourse-level representations, and making pragmatic inferences. A major goal of both behavioral psycholinguistics and cognitive neuroscience of language is to understand which cognitive mechanisms support language comprehension, and whether and how these mechanisms are shared with other (nonlinguistic) cognitive functions.

Psycholinguists have long invoked "domain-general constructs" when discussing lexical access and syntactic/semantic dependency formation, from storage and retrieval of information from working memory, to updating focal attention, inhibiting irrelevant information, selecting an option among alternatives, and predictive processing (Johnson-Laird 1983; Abney and Johnson 1991; King and Just 1991; Resnik 1992; Gernsbacher 1993; Gibson 1998, 2000; McElree 2000, 2001; Gordon et al. 2002; Lewis and Vasishth 2005; Fedorenko et al. 2006, 2007; Lewis et al. 2006; Novick et al. 2009; Rodd et al. 2010; Schuler et al. 2010; Vergauwe et al. 2010; van Schijndel et al. 2013; Rasmussen and Schuler 2018; inter alia). If some linguistic processes require these or other domain-general operations, does it mean that language shares neural mechanisms with other domains?

It has long been known that language processing recruits particular neural circuitry (Broca 1861; Wernicke 1874; Geschwind 1970). However, prior cognitive neuroscience work has argued both 1) that some of this circuitry (e.g., "Broca's area") may not be specialized for language processing per se, but rather used for broader cognitive functions—like hierarchical syntactic structure building—that operate not only in language but also in other domains like music, mathematics, and action planning (Patel 2003, 2012; Tettamanti and Weniger 2006; Fadiga et al. 2009; Friedrich and Friederici 2009; Slevc et al. 2009; Anderson 2010; Fitch and Martins 2014; Rodriguez and Granger 2016; inter alia, see Fedorenko and Blank 2020 for a review); and 2) that language processing relies on a more spatially distributed network, extending beyond the "classic" language areas, that includes regions traditionally associated with domain-general executive control (Mesulam 1998; Kaan and Swaab 2002; Kuperberg et al. 2003; Novick et al. 2005; Rodd et al. 2005; Thompson-Schill et al. 2005; Novais-Santos et al. 2007; January et al. 2009; Peelle et al. 2010; Rogalsky and Hickok 2011; McMillan et al. 2012, 2013; Wild et al. 2012; Blumstein and Amso 2013; Nieuwland et al. 2012; Hsu and Novick 2016; Hagoort 2019; inter alia). Hypotheses from psycholinguistics, cognitive science, and cognitive neuroscience therefore converge to predict a role for domain-general executive resources in human language comprehension.

Within the human brain, the most plausible place to look for domain-general recruitment is in the fronto-parietal/cingulo-opercular "multiple demand (MD)" network, which supports a broad range of executive functions, including inhibitory control, attentional selection, conflict resolution, and maintenance and manipulation of task sets (Duncan 2010, 2013; Fedorenko et al. 2013; Duncan et al. 2020; Assem et al. 2020b). Indeed, MD regions have been shown to be sensitive to linguistic processing difficulty (e.g., due to ambiguity or complexity in unambiguous structures) across diverse manipulations (Kuperberg et al. 2003; Rodd et al. 2005; Novais-Santos et al. 2007; January et al. 2009; Peelle et al. 2010; Nieuwland et al. 2012; McMillan et al. 2013; inter alia). Further, activity in this network has been shown to correlate positively with reaction times—a behavioral measure of processing difficulty—across tasks (Yarkoni et al. 2009; Taylor et al. 2014). If indeed MD regions register processing load during language comprehension, this would support the hypothesis that domain-general resources are engaged in language comprehension.

The ability of prior work to bear on this hypothesis is limited by 2 factors. First, language comprehension effort has typically been studied by relating theory-driven linguistic variables (e.g., word frequency, word predictability, structural complexity, constituent length, etc.) to neural activity (Mazoyer et al. 1993; Stowe et al. 1998; Vandenberghe et al. 2002; Friederici et al. 2003; Dronkers et al. 2004; Humphries et al. 2006; Brennan et al. 2010; Pallier et al. 2011; Rogalsky and Hickok 2011; Brennan and Pylkkänen 2012; Brennan et al. 2016; Henderson et al. 2016; Willems et al. 2016; Lopopolo et al. 2017; Nelson et al. 2017; Shain et al. 2019; inter alia). Despite the critical role of theory in understanding human cognition, theory-driven variables are only as good as the underlying theory and can only be expected to capture a fraction of the language comprehension effort given the multifaceted nature of language. Such variables may fail to characterize some components of language comprehension and thereby underestimate the extent to which some neural circuits, including the domain-general MD circuits, are implicated in comprehension. And second, prior work, including many of the aforementioned studies purportedly showing MD involvement in language comprehension, has generally relied on language stimuli cleverly constructed to directly manipulate some aspect of language processing difficulty and has often included tasks on top of language comprehension, like making judgments about sentences or deciding whether a sentence matches a picture (Friederici et al. 2003; Fiebach et al. 2004; Rodd et al. 2005; Bilenko et al. 2008; Kuperberg et al. 2008; Snijders et al. 2009). Such artificial hand-constructed stimuli and the presence of extraneous tasks (i.e., tasks other than language comprehension per se) make language processing in these studies very different from natural comprehension "in the wild," and may inadvertently trigger recruitment of domain-general problem solving and task strategizing mechanisms that would not be at play in natural-comprehension scenarios (Small and Nusbaum 2004; Hasson and Honey 2012; Campbell and Tyler 2018; Hasson et al. 2018; Diachek et al. 2020). Such stimuli and tasks might thus overestimate MD involvement in language comprehension, especially given the sensitivity of MD regions to task demands (Miller and Cohen 2001; Sreenivasan et al. 2014; D'Esposito and Postle 2015). MD recruitment for language processing would therefore be better supported if an MD response to theory-neutral measures of comprehension difficulty could be shown under more naturalistic experimental conditions.

Several recent neuroimaging studies have used naturalistic language stimuli (see Supplementary Table 1). However, to our knowledge, only one study (Henderson et al. 2015) has attempted to relate brain activity to reading latencies using naturalistic materials. Henderson et al. (2015) used eye-tracking (ET) during reading in functional magnetic resonance imaging (fMRI) and found that blood oxygen level–dependent (BOLD) activity in parts of the left middle/superior temporal gyri scaled with fixation duration: stronger responses to words that are fixated for longer during reading. This relationship was more pronounced during the reading of meaningful texts compared with texts devoid of meaning (generated by replacing each letter in a text by a geometric shape), suggesting that longer fixations were partially driven by effort to update linguistic representations. Crucially, this result was obtained by regressing participants' moment-to-moment brain activity against their own word-by-word reading latencies, leaving open the possibility that the results are confounded by nonlinguistic, stimulus-independent tertiary variables like attentional fluctuations and saccade planning.

Building on Henderson et al. (2015), and to directly test the hypothesis of domain-general executive involvement in language comprehension, we use context-rich, naturalistic language stimuli presented auditorily without any extraneous tasks and correlate 1) experimentally-obtained behavioral reaction time measures of language processing difficulty during reading,

with 2) fMRI measures of activity in the domain-general MD network. To increase the interpretability of such correlations, we compare them with brain-behavior correlations based on a different functional network: the domain-specific, fronto-temporal "core language network" (e.g., Binder 1997; Jung-Beeman 2005; Fedorenko and Thompson-Schill 2014). This network serves as a good comparison for the MD network because it robustly engages in comprehension (during both listening and reading; Fedorenko et al. 2010; Regev et al. 2013; Scott et al. 2017; Deniz et al. 2019) but shows little to no engagement in other high-level cognitive processes (Ivanova et al. in prep.; Monti et al. 2009, 2012; Fedorenko et al. 2011; Pritchett et al. 2018; for reviews see (Fedorenko and Blank 2020; Fedorenko and Varley 2016). Below, we describe and justify the main design features of our experiment.

Our use of behavioral reading data as a global proxy for comprehension difficulty follows a standard psycholinguistic approach where reaction times are examined for linguistic materials whose comprehension requires different kinds of (hypothesized) computations, in either experimentally constructed materials (e.g., Frazier and Rayner 1987; Clifton and Frazier 1989; Gibson 1991, 1998; Grodner et al. 2002; Levy 2008), or naturalistic ones (e.g., Demberg and Keller 2008; Smith and Levy 2013). Although incremental reading data are known to have a complex relationship to mental states (Posner 1980, 2016; Remington 1980; Klein and Farrell 1989; Wright and Ward 2008; inter alia) and to be sensitive to nonlinguistic factors like general attention, sensory/perceptual processing, motor control, and task-related strategizing (Rayner 1998; Kennedy 2000; Kaakinen and Hyönä 2010; Schotter et al. 2014), a premise underlying most psycholinguistic work in this domain is that incremental behavioral measures of reading effort track language-related comprehension difficulty with sufficient reliability such that they can be used to evaluate theories of human sentence comprehension (Rayner 1977, 1978, 1998; Just and Carpenter 1980; Mitchell 1984; Lewis et al. 2006). Furthermore, our experimental design reduces the influence of idiosyncratic processes such as attention fluctuations by 1) aggregating reading data from many participants; 2) separating the samples that provide behavioral data from the sample providing the neuroimaging data; and 3) using different presentation modalities across the behavioral (visual) and fMRI (auditory) paradigms (cf. Henderson et al. 2015).

To elaborate, when behavioral measures and fMRI signal are acquired from the same participant, many potential sources of covariation may exist between these signals that are not directly linked to language processing and that we would ideally like to factor out, including changes in the level of fatigue and general attention. Instead, we are interested in cognitive demands linked to changes in the "linguistic stimulus properties." Measuring behavioral comprehension difficulty in one cohort and using those measures to predict the fMRI activity of another cohort should capture these stimulus-related processes. Furthermore, because our design uses comprehension difficulty measures from reading to predict fMRI activity during listening, it minimizes potential contributions from perceptual (visual/auditory) processing (such as effects related to speaker word rate), and thus helps isolate modality-independent "comprehension demands" related to word-level and combinatorial (semantic and syntactic) processing, as well as higher-level discourse processing. Demands related to word-level processing have to do with how easy each word is to access from memory (which is known to depend on factors like word length, frequency, etc. (Howes and Solomon 1951; Hudson and Bergman 1985; Pickering and Frisson 2001; De Deyne et al. 2013); demands related to combinatorial processing have to do with the length of inter-word dependencies, how expected a particular structure or word sequence is, whether a temporary or global ambiguity is present, etc. (Dopkins et al. 1992; Sturt 2007; Smith and Levy 2013); and demands related to discourse-level processing have to do with the ease of establishing coherence across clauses and sentences (Gordon and Chan 1995; Gordon and Scearce 1995; Chambers and Smyth 1998). In summary, the current design is intended to "distill stimulus-related, generalizable variation in comprehension difficulty."

We consider 2 different behavioral methods—self-paced reading (SPR, Aaronson and Scarborough 1977; Just et al. 1982) and ET during reading (Rayner 1998), from 2 large, existing datasets (Futrell et al. 2018, 2020; von der Malsburg et al. unpublished). These measures of comprehension effort serve as theory-neutral, broad-coverage estimates of computational load during language comprehension, since they should permit detection of any mechanisms that contribute to processing latencies, even if their role is not yet captured by any existing theory.

When correlating these measures with neuroimaging data, we consider the detailed time-course of activation during listening, rather than aggregate measures averaging across parts of the stimulus. The time-varying fMRI data enable us to exploit relatively fine-grained variation in incremental processing difficulty that may be attenuated in aggregate measures. In addition, we infer the hemodynamic response from the data directly, in order to account for individual and regional variation in the underlying hemodynamic response (Handwerker et al. 2004). Finally, we employ nonparametric hypothesis tests on out-of-sample data, in order to increase the statistical robustness of the results and reduce the risk of replication failure (Menke and Martinez 2004; Eklund et al. 2012).

To foreshadow our results, we find that reading latencies predict neural activity in the core language network, but not in the MD network. This finding supports the hypothesis that incremental processing effort during naturalistic language comprehension is largely restricted to neural circuits–and, by extension, cognitive resources–that are specialized for language comprehension, with little role played by domain-general executive systems.

Materials and Methods

Short Stories

We use the Natural Stories Corpus (Futrell et al. 2018, 2020; data downloaded from https://github.com/languageMIT/naturalstories.git), which contains 10 stories that were constructed from existing, publicly available texts (fairy tales, short stories, and Wikipedia articles) but edited so as to make comprehension difficulty more variable than in fully natural texts. The dataset includes recordings of these stories by 2 native English speakers (one male, E.G., and one female).

Behavioral SPR Data

The Natural Stories Corpus includes SPR data from 181 native English-speaking participants recruited through Amazon.com's Mechanical Turk. Participants gave informed consent in accordance with the Internal Review Board at the Massachusetts

Institute of Technology (MIT) and were paid for their participation. Participants read stories in a moving-window self-paced word-by-word reading paradigm, where a button has to be pressed to reveal each subsequent word. The time spent on each word provides an overall estimate of processing difficulty at that point in the sentence/story. Each story was followed by 6 multiple-choice comprehension questions and if a participant answered fewer than 5 questions correctly, their reading time data for that story were excluded. Outlier reading times of < 100 ms or > 3000 ms were also excluded. These exclusion criteria were the ones followed by Futrell et al. (2018). Reading times were aggregated across participants for each word, and the average word reading time across stories was 335 ms. As a result, for each word in each story, we have a single (average) reading time.

Behavioral ET Study

In total, 40 native English-speaking participants recruited from the University of California, San Diego (UCSD) undergraduate population gave informed consent in accordance with the Internal Review Board at UCSD and were paid for their participation. They read the stories in an ET paradigm. A tower-mounted EyeLink 1000 eye-tracker recorded eye movements as participants read the stories presented a few sentences at a time (the boundaries among the story fragments and lines within fragments differed across participants so as to vary the words that span the screen-change and line boundaries). Each story was followed by 2 true/false comprehension questions. Software for automatic correction of eye fixations was used to repair data recorded with imperfect eye-tracker calibration (Cohen 2013). A set of heuristics were used to detect and remove episodes of track loss, poor-quality data, and episodes where reader merely skimmed the text. In particular, fixations were removed when 1) the previous and/or subsequent fixations were 5 or more words away which is indicative of skimming (all the skipped words were also removed from the subject's data in this case), 2) initial fixations on a new page of text occurred on words that were not at the beginning of the text, 3) the fixations could not be mapped to any word, or 4) consecutive fixations were moved in different directions by Cohen correction (Cohen et al. 2013).[1] For each word, 4 canonical ET measures were calculated (first pass regression, regression path duration, first pass reading time, and first fixation progressive) which are believed to index different linguistic processes involved in reading, ranging from word recognition to high-level discourse integration (Rayner 1998; Vasishth et al. 2013). ET measures were aggregated across participants for each word. As a result, for each word in each story, we have 4 (average) ET measures.

fMRI Experiment

Participants

In total, 42 right-handed native English speakers (average age 22.7, standard deviation [SD] = 3.3; 24 females) from the MIT community gave informed consent in accordance with the Internal Review Board at MIT and were paid for their participation. (Subsets of this dataset were used by Blank et al. 2014; Blank and Fedorenko 2017; Shain et al. 2020).

General Approach

Each participant listened to a subset of the stories from Futrell et al. (2018) and performed one or more "localizer" tasks (e.g., Saxe et al. 2006) used to identify the brain networks of interest.

Critical Task. Participants listened to the recordings of the spoken stories. Each story corresponded to one fMRI run. Eight of the 10 stories were used, and any given participant heard between 2 and 8 stories (average = 4; 2 stories: $n = 12$, 3 stories: $n = 13$, 4 stories: $n = 2$, 5 stories: $n = 4$, 6 stories: $n = 5$, 7 stories: $n = 1$, 8 stories, $n = 5$). Each story lasted between 4.5 and 6 min. Participants were asked to listen attentively. At the end of each story, a set of six 2-alternative forced-choice comprehension questions appeared one by one, and participants answered by pressing 1 of 2 buttons. These questions were designed to be challenging and required attentive listening and the ability to respond quickly. On average, participants failed to provide an answer to 11.5% of the questions (SD = 15.2%) and, on the remaining questions, their mean accuracy was 83.5% (SD = 10.1%). (Comprehension data were available for 33 participants: they were lost for 2 participants, not recorded for 3 participants due to a script error, and not collected for 4 participants who listened to the stories as part of a larger experiment for which the design did not include comprehension questions). A binomial test for each participant (uncorrected across participants) showed that all but 1 participant (of those for whom behavioral data was available) demonstrated above-chance accuracy ($P < 0.01$). In the supplementary materials, we report our main analysis restricted to participants with very good performance, which revealed the same general pattern of results (compare Fig. 3*b* and Supplementary Fig. 3.).

Localizer Tasks. Most participants also performed 2 localizer tasks. These tasks were used to functionally identify the 2 networks of interest: the MD network, and the language network. To identify the MD regions, we used a visuo-spatial working memory task that included a hard and an easy condition (Fedorenko et al. 2011; Assem et al. 2020a). In both conditions, participants kept track of sequences of locations presented in a 3×4 grid. In the hard condition, 8 locations were presented, 2 at a time; in the easy condition, 4 locations were presented, one at a time (for timing details, see Fig. 1*a* in Assem et al. 2020a). At the end of each trial, participants performed a 2-alternative forced-choice task to indicate the set of locations they had just seen. We used a standard blocked design (with 4 trials per block), with condition order counterbalanced across runs. Each participant completed 2 runs. The *Hard* > *Easy* contrast targets brain regions that support executive functions, like attention, working memory, and cognitive control (Duncan and Owen 2000; Duncan 2010, 2013). This contrast has been previously found to robustly activate MD regions (Fedorenko et al. 2013; Blank et al. 2014; Assem et al. 2020a), which have been shown to respond to difficulty manipulations across diverse tasks (Duncan and Owen 2000; Fedorenko et al. 2013; Hugdahl et al. 2015; Shashidhara et al. 2019).

Because only 35 of the 42 participants had completed the working memory localizer, for an alternative analysis we used another way to identify MD regions. In particular, we used the *Nonwords* > *Sentences* contrast from the language localizer task, described below. This contrast has been previously validated as a robust MD network localizer: it can reliably localize MD regions bilaterally, and generalizes across a wide array of stimuli

1 The Cohen correction is designed to correct for poor eye-tracker calibration. However, poor calibration is reflected in fixation offsets in the same direction, and variable correction vectors therefore indicate that the Cohen correction failed.

and tasks, both linguistic and nonlinguistic (Fedorenko et al. 2013). The nonwords condition is associated with greater processing effort and elicits a response stronger than the sentences condition both in the version with a memory probe and a passive reading version with a button press (e.g., Mineroff et al. 2018). We verified that the MD regions localized with the *Nonwords* > *Sentences* contrast show a robust *Hard* > *Easy* effect in the visuo-spatial working memory task in our data for the subset of 35 participants who have done both localizers. In particular, all MD regions showed a stronger response to the hard condition than the easy condition (dependent samples $t_{(35)} > 3.84$, $P < 10^{-6}$, false discovery rate corrected for the number of regions; Cohen's $d > 0.30$, computed based on an independent samples formula, see Supplementary Fig. 1). Both ways of defining the MD ROIs led to similar results for the critical brain-behavior analysis (compare the results for ROIs defined with the *Hard* > *Easy* contrast in the visuo-spatial working memory task in Fig. 3 and those for ROIs defined with the *Nonwords* > *Sentences* contrast in Supplementary Fig. 2), as described in Table 2.

To identify the language regions, we used a reading task that included sentences and lists of unconnected, pronounceable nonwords, as described in detail in Fedorenko et al. (2010). In both conditions, participants read the stimuli, presented one word/nonword at a time (for timing parameters, see Table 1). Eighteen participants read the materials passively (a button-press task at the end of each trial was included in order to maintain alertness); for the remaining 24 participants, each trial ended with a memory probe, that is, a word/nonword, and they had to indicate (via a button press) whether or not this probe had appeared in the preceding sentence/nonword sequence. We used a standard blocked design (with 3–5 trials per block; Table 1), with condition order counterbalanced across runs. Each participant completed 2–4 runs of the localizer task. (A version of this localizer is available from https://evlab.mit.edu/funcloc/download-paradigms.) The *Sentences* > *Nonwords* localizer contrast targets brain regions that support high-level language comprehension. This contrast generalizes across tasks (Fedorenko et al. 2010; Regev et al. 2013; Scott et al. 2017) and presentation modalities (reading vs. listening; e.g., Fedorenko et al. 2010; Braze et al. 2011; Vagharchakian et al. 2012; Scott et al. 2017; Deniz et al. 2019). All the regions identified by this contrast show sensitivity to the processing of word meanings (e.g., stronger responses to real words than nonwords) and combinatorial syntactic and semantic processing (e.g., stronger responses to sentences and Jabberwocky sentences than to unstructured word and nonword sequences) (Keller et al. 2001; Rodd et al. 2005; Heim et al. 2008; Fedorenko et al. 2010, 2012; 2016, 2020a; Bautista and Wilson 2016; Blank et al. 2016; Mineroff et al. 2018; Mollica et al. 2020). The *Sentences* > *Nonwords* contrast encompasses all of these processes, but narrower contrasts that target a subset of them identify the same cortical network (e.g., Fedorenko et al. 2010), suggesting that all the regions in the fronto-temporal language network support all of these high-level linguistic processes (for discussion, see Fedorenko et al. 2020a, 2020b). In addition, the same network is identified by broader contrasts that do not subtract out phonological processing and also include pragmatic and discourse-level processes (e.g., a contrast between natural spoken paragraphs and their acoustically degraded versions or paragraphs in an unfamiliar language (Ayyash et al. in prep.; Ivanova et al. in prep.; Scott et al. 2017). Finally, this localizer also identifies right-hemisphere homotopic areas of the classic, left-hemisphere language areas (e.g., Mahowald and Fedorenko 2016), which we included here because our other network of interest (the MD network) is bilateral and because right-hemisphere language regions have been previously implicated in several aspects of language comprehension (Jung-Beeman 2005; Wehbe et al. 2014; Huth et al. 2016; Deniz et al. 2019).

fMRI Data Acquisition

Structural and functional data were collected on the whole-body 3-Tesla Siemens Trio scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted structural images were collected in 176 sagittal slices (1-mm isotropic voxels; repetition time [TR]: 2530 ms; echo time [TE]: 3.48 ms). Functional BOLD data were acquired using an echo planar imaging sequence with a flip angle of 90° and applying generalized autocalibrating partially parallel acquisition with an acceleration factor of 2. Images were collected in 31 near-axial slices, acquired in an interleaved order with a 10% distance factor (in-plane resolution: 2.1 × 2.1 mm; slice thickness: 4 mm; field of view: 200 mm in the phase encoding anterior to posterior [A >> P] direction; matrix size: 96 × 96; TR: 2000 ms; TE: 30 ms). Prospective acquisition correction (Thesen et al. 2000) was used to adjust the positions of the gradients based on the subject's head motion one TR back. The first 10 s of each run was excluded to allow for steady-state magnetization.

fMRI Data Preprocessing

Spatial Preprocessing

Data preprocessing was carried out with SPM5 and custom MATLAB scripts. (We used an older version of SPM because data for this study are used across other projects spanning many years and hundreds of participants, and we wanted to keep the SPM version the same across all the participants.) Preprocessing of anatomical data included normalization into a common space (Montreal Neurological Institute [MNI]) template, resampling into 2-mm isotropic voxels, and segmentation into probabilistic maps of the gray matter, white matter (WM), and cerebrospinal fluid (CSF). Preprocessing of functional data included motion correction, normalization, resampling into 2-mm isotropic voxels, smoothing with a 4-mm FWHM Gaussian kernel, and high-pass filtering at 200 s.

Temporal Preprocessing

Additional preprocessing of data from the story comprehension runs was carried out using the CONN toolbox (Whitfield-Gabrieli and Nieto-Castanon 2012) with default parameters, unless specified otherwise. Five temporal principal components of the BOLD signal time-courses extracted from the WM were regressed out of each voxel's time-course; signal originating in the CSF was similarly regressed out. Six principal components corresponding to the 6 motion parameters estimated during offline motion correction were also regressed out, as well as their first time derivative. No low-pass filtering was applied.

Modeling Localizer Data

For each localizer task, a general linear model estimated the effect size of each condition in each experimental run in each voxel. These effects were each modeled with a boxcar function (representing entire blocks) convolved with the canonical hemodynamic response function (HRF). The model also included first-order temporal derivatives of these effects, as well as nuisance

Table 1 Summary of the procedural and timing details for the different versions of the language localizer used in the current study

	Version			
	I	II	III	IV
Number of participants	24	7	6	5
Task: passive reading/memory probe?	PR	MP	MP	MP
Conditions	Sentences, Nonwords	Sentences, Word lists, Nonwords	Sentences, Nonwords	Sentences, Word lists, Nonwords
Words/nonwords per trial	12	12	12	8
Trial duration (ms)	6000	6000	6000	4800
Fixation	100	300	300	300
Presentation of each word/nonword	450	350	350	350
Memory probe	—	1000	1000	1350
Fixation	500	500	500	350
Trials per block	3	3	3	5
Block duration (s)	18	18	18	24
Blocks per condition per run	8	6	8	4
Fixation block duration (s)	14	18	18	16
Number of fixation blocks per run	5	4	5	3
Total run time (s)	358	396	378	336
Number of runs	2	2–3	2	3–4

regressors representing entire experimental runs and offline-estimated motion parameters. The obtained beta weights were then used to compute the functional contrast of interest: *Nonwords* > *Sentences* for the MD localizer, and *Sentences* > *Nonwords* for the language localizer.

Defining Functional Regions of Interest

For each participant, functional regions of interest (fROIs) were defined by combining 2 sources of information (following Fedorenko et al. 2010; Julian et al. 2012): 1) the participant's activation map for the relevant localizer contrast, and 2) group-level spatial constraints ("masks"). The latter demarcated brain areas within which most or all individuals in prior studies showed activity for the localizer contrasts (Fig. 1).

For the MD fROIs, we used masks derived from a group-level probabilistic representation of data from the *Hard* > *Easy* contrast from the visuo-spatial working memory MD-localizer task in a set of 197 participants. These masks were constrained to be bilaterally symmetric by averaging individual contrast maps across the 2 hemispheres prior to generating the group-level representation. The topography of these masks (available for download from http://web.mit.edu/evelina9/www/funcloc/funcloc_parcels.html) largely overlapped with anatomically based masks that were used in some prior studies (e.g., Fedorenko et al. 2013; Blank et al. 2014; Paunov et al. 2019). In particular, 10 masks were used in each hemisphere: in the posterior (PostPar), middle (MidPar), and anterior (AntPar) parietal cortex, precentral gyrus (PrecG), superior frontal gyrus (SFG), middle frontal gyrus (MFG) and its orbital part (MFGorb), opercular part of the inferior frontal gyrus (IFGop), the anterior cingulate cortex and presupplementary motor cortex (ACC/pSMA), and the insula (Insula).

For the language fROIs, we used masks derived from a group-level probabilistic representation of data from the *Sentences* > *Nonwords* contrast from the language localizer task in a set of 220 participants. These masks (available for download from http://web.mit.edu/evelina9/www/funcloc/funcloc_parcels.html) were similar to the masks derived from 25 participants, as originally reported in Fedorenko et al. (2010), and covered extensive portions of the left lateral frontal and temporal cortex. In particular, 6 masks were used: 3 in the frontal lobe (in the inferior frontal gyrus [IFG], and its orbital part [IFGorb], and in the middle frontal gyrus [MFG]), and 3 in the temporal and parietal cortex (in the anterior temporal cortex [AntTemp], posterior temporal cortex [PostTemp], and in the angular gyrus [AngG]). Note that although both the MD and the language parcel sets include a parcel within the opercular IFG (called "IFGop" in both sets), the MD and language fROIs within this parcel are largely nonoverlapping within any given individual (see e.g., Blank et al. 2014 and Paunov et al. 2019 for quantification of overlap; see Fedorenko and Blank 2020, for discussion). The left hemisphere (LH) masks were mirror-projected onto the RH to create 6 homotopic masks, to define the RH language fROIs.

For more detail (see also Fedorenko et al. 2010), each set of masks was generated as follows: 1) for each individual in a large sample ($n = 197$ for the MD masks, and $n = 220$ for the language masks), a lenient statistical whole-brain threshold ($P < 0.001$, uncorrected) for the relevant contrast was used to generate a binary contrast map, where 1 = a significant effect, and 0 = otherwise; 2) the individual thresholded maps are overlaid in the common space to create a group-level probabilistic overlap map where, for each voxel, the percentage of participants showing a significant effect is represented; 3) following smoothing and removing voxels where fewer than 10% of participants show the effect, this probabilistic map is divided into regions (masks) using a watershed algorithm, and the subset of masks where a large fraction of participants show activation and where the target contrast is replicable across runs is selected. This is done once, and then the same set of masks is used in all future studies that use the same or similar localizer. (There is no reason to create the masks for each new dataset using the localizer data form *that* set of participants because for robust

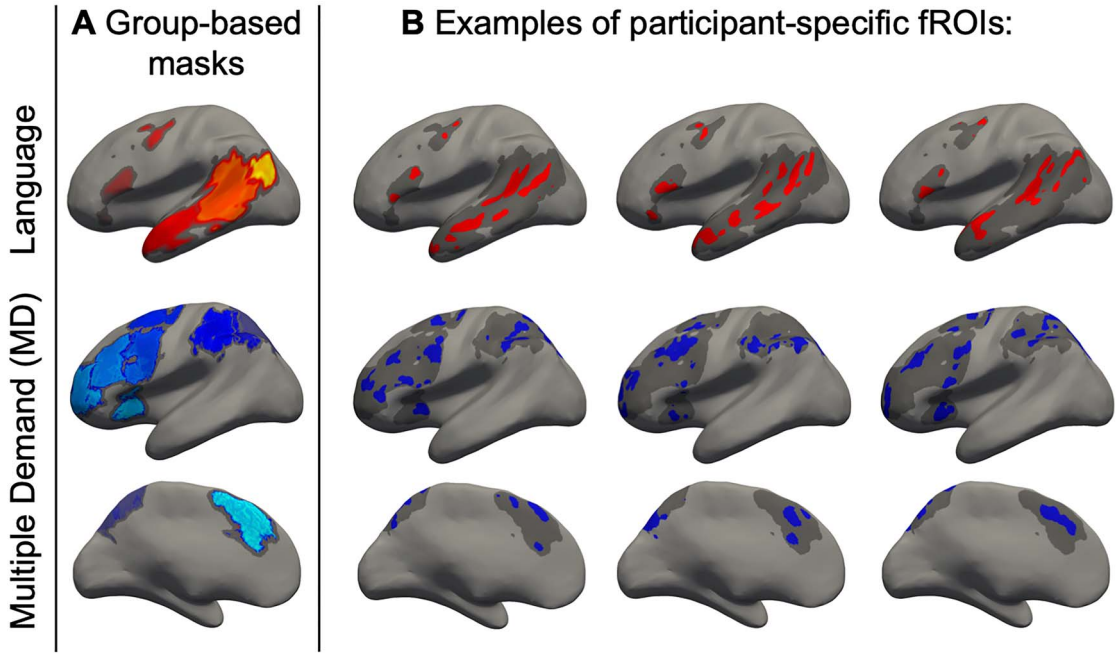

Figure 1. Defining participant-specific fROIs in the language (top) and MD (bottom) networks (only the left-hemisphere is shown, the MD network is defined using the visuo-spatial *Hard* > *Easy* contrast). All images show approximated projections from functional volumes onto the surface of an inflated brain in common space. (*a*) Group-based masks used to constrain the location of fROIs. (*b*) Example fROIs of 3 participants. Note that, because data were analyzed in volume (not surface) form, some parts of a given fROI that appear discontinuous in the figure (e.g., separated by a sulcus) are contiguous in volumetric space.

contrasts, like those typically used as localizer contrasts, similar parcels emerge once you reach ~20–30 participants; besides, these masks are, by design, sufficiently large to encompass inter-individual variability in the precise locations of the functional areas. The advantage of these "functional" group-level masks over for example, masks based on macro-anatomical areas is that these regions better respect the activation landscape given that in many cases, activations will straddle anatomical boundaries.) In any new experiment, like the current study, each set of masks is intersected with the individual activation maps for the corresponding contrast (e.g., the MD masks are intersected with the *Hard* > *Easy* contrast), and in each participant within each mask, the most responsive voxels are selected as the fROI.

In particular, for each participant, 20 MD fROIs were created by intersecting each MD mask with that participant's unthresholded *t*-map for the relevant contrast (the *Hard* > *Easy* contrast in the visuo-spatial working memory task for the main analysis, and the *Nonwords* > *Sentences* contrast for the supplementary analysis), and the 10% of voxels with the highest *t*-values within each mask were chosen as the fROI. In a parallel fashion, 12 language fROIs were created by intersecting each language mask with that participant's unthresholded *t*-map for the *Sentences* > *Nonwords* contrast, and the 10% of voxels with the highest *t*-values within each mask were chosen as the fROI. Finally, a BOLD signal time-course for each story in the critical story listening task was extracted from each voxel in each fROI of each participant.

We note that, for both the MD and language networks, only the masks—which cover large swaths of cortex—were symmetric across hemispheres; the fROIs themselves were free to vary in their location between the 2 hemispheres, within the borders of these masks.

Model of Comprehension Difficulty Using SPR Times

As noted above, a separate cohort of participants ($n = 181$) was used for this study. To verify that SPR times (SPRTs) reflect stimulus-related processing (following the logic in Hasson et al. 2008; Lerner et al. 2011), we first computed inter-subject correlations among the time-series of per-word SPRTs for each story: each individual's time-series was correlated with the average time-series of the rest of the participants. The average correlations varied between $r = 0.38$ and $r = 0.59$ across the stories and were all reliably above chance (all $Ps < 10^{-25}$). As mentioned above, we used the default exclusion criteria used by (Futrell et al. 2018): we excluded data for a story if a participant answered < 5 out of 6 questions wrong and outlier reading times of < 100 ms or > 3000 ms were also excluded.

Mean reading times per word were temporally aligned with their corresponding word onsets in the auditory stimulus. Then, we obtained a per-TR time-series of SPRTs by averaging the reading times for the words that occurred within each TR (corresponding to 2 s) when the recorded stories were played in the fMRI experiment. Words that overlapped with 2 TRs were assigned to the TR with greater overlap. We then computed the average (across participants) per-TR SPRT, arriving at a final measure of comprehension difficulty at each TR.

Model of Comprehension Difficulty Using ET Measures

As noted above, a separate cohort of participants ($n = 40$) was used for this study. We used 4 ET measures for participants in the ET study ($n = 40$): 1) first pass regression (FPR), a variable indicating for each word whether or not a regressive eye-movement occurred from that word in the first pass; 2)

regression path duration (RPD) or go-past time, the duration of the period between the onset of the first fixation on a word and the first fixation on anything to the right of it (RPD thus includes time spent on regressive fixations); 3) first pass reading time (FPRT) or gaze duration, the summed duration of all first-pass fixation durations on a word before any other word (left or right) is fixated; and 4) first fixation progressive (FFP), a variable indicating whether the first fixation on a word took place before any downstream words were viewed. To verify that ET measures reflect stimulus-related processing, we followed the same procedure as used for SPRTs, and compute inter-subject correlations among the time-series of per-word FPRs, RPDs, FPRTs, and FFPs for each story. The average correlations varied between $r = 0.13$ and $r = 0.17$ across the stories for FPRs, between $r = 0.27$ and $r = 0.42$ across the stories for RPDs, between $r = 0.38$ and $r = 0.53$ across the stories for FPRTs, and between $r = 0.37$ and $r = 0.53$ across the stories for FFPs (all correlations higher than chance, all $Ps < 10^{-4}$).

Mean ET measures per word were temporally aligned with their corresponding word onsets in the auditory stimulus. Then, we obtained a per-TR average measure of FPR, RPD, FPRT, and FFP by averaging each of the 4 ET measures across the words that occurred within each 2 s TR, and then averaging these values across participants, following the same procedure as used for SPRTs.

Critical Analysis Using SPR Times and ET Measures

Our analysis is summarized in Fig. 2. As described above, for each TR t, we obtained SPRT, FPR, RPD, FPRT, and FFP measures. We constructed a design matrix for the experiment in which each row t corresponds to the concatenated 5 measures for a TR t. To account for the hemodynamic response, we modeled its effect as a fourth order finite impulse response (FIR) filter. We performed a simple linear regression: for each of the 5 variables, we estimate 4 weights that correspond to TRs $t+1$, $t+2$, $t+3$, and $t+4$ after onset at time t. Effectively, this corresponds to concatenating delayed versions of the design matrix so that each row t in the final design matrix contains the 5 measures for TRs $t-4$, $t-3$, $t-2$, and $t-1$. This is a common approach for accounting for the hemodynamic response delay (Wehbe et al. 2014; Huth et al. 2016), and the choice of an 8 s window is typically used to capture the effect of stimulus features on the fMRI response. This encoding model analysis (Wehbe et al. 2014; Huth et al. 2016) differs from the typical GLM analysis in 2 ways. First, instead of assuming a fixed HRF that is constant across the brain, this approach allows for variability in the hemodynamic response by implicitly estimating it at each voxel. And second, instead of running a significance test on the regression weights, we run a more stringent test: we assess the generalization and stability of the learned weights by using them to predict held-out fMRI data unseen in training.

In particular, for each participant, we estimated generalization via a cross-validation scheme in which we iteratively held out one story and learned the regression weights from the remaining stories. We then predicted BOLD activity for the held-out story using the (delayed) design matrix for that story and the learned regression weights. This procedure resulted in a predicted time-series of BOLD activity in each voxel in each fROI of each participant during the held-out story. We then measured how closely these predictions correspond to the real data via Pearson's correlation. This correlation was computed between the average (across voxels) fROI activity predicted by the model and the corresponding average fROI activity in the real data to obtain summary statistics for the fROIs. Finally, we averaged the correlation values across all cross-validation folds to obtain a single correlation value per fROI per participant (i.e., 35 mean correlation values for each fROI, one for each participant when considering only the participants with the visuo-spatial MD localizer and 42 mean correlation values when considering all participants).

To better characterize the findings at the level of the networks of interest, the above analysis was repeated, but this time, predicted and actual BOLD time series were grouped into 4 sets: Left Hemisphere (LH) Language fROIs, Right Hemisphere (RH) Language fROIs, LH MD fROIs, and RH MD fROIs.

It is worth mentioning that the direction of prediction we used here (predicting brain activity from comprehension difficulty measures instead of the other way around) was in keeping with the encoding model approach (Naselaris et al. 2011; Wehbe et al. 2014; Huth et al. 2016) and does not imply that brain activity is caused by comprehension difficulty measures. Typically, the use of encoding models where stimulus features are used to predict brain activity allows researchers to make statements about a particular stimulus causing neural responses (Weichwald et al. 2015). However, here we use the encoding approach only as a way to test the relationship between 2 measures (fMRI activity and reading times) related to the same cause (comprehension difficulty). This setup does not therefore have the causal implications of the usual encoding analyses, nor do we attempt to make any causal claims.

Noise Ceiling Correction

To help with interpreting prediction performance, we provide measures of prediction performance that are corrected by the estimated noise ceiling for each fROI and fROI group (we provide raw prediction performance measures in Supplemental). The noise ceiling is an approximation of the maximum possible performance. fMRI stimuli engage brain regions to a different extent, and regions have different physiological characteristics, both of which affect the signal-to-noise ratio. We estimate the noise ceiling for each fROI (and fROI group) across participants by adapting the method proposed by Hsu et al. (2004) to be used for multiple participants. To evaluate the noise ceiling, Hsu et al. (2004) consider different repeats of the same stimulus that is seen by multiple participants. We treat the average fROI activity for the subjects listening to the same story as different repeats of the same story. For each story and each fROI, we evaluate the noise ceiling by first computing the average time-course of this fROI for each of the k subjects that have listened to this story. We then compute the correlation of each of the $\binom{k}{2}$ pairs of time series. We then average all these pairwise correlations, and further average these estimates for all the stories. We end up with a measure of noise ceiling for each fROI. We repeat this approach for fROI groups. Following previous work (Hsu et al. 2004; Lescroart et al. 2015; Lescroart and Gallant 2019), we normalize the average prediction performance by the square-root of the noise-ceiling, yielding normalized correlation values.

Computing Confidence Intervals

The participant-specific (unnormalized and normalized) correlation values were averaged across participants, and empirical confidence intervals were estimated for the mean prediction in each fROI, by running a bootstrap test that takes into account the

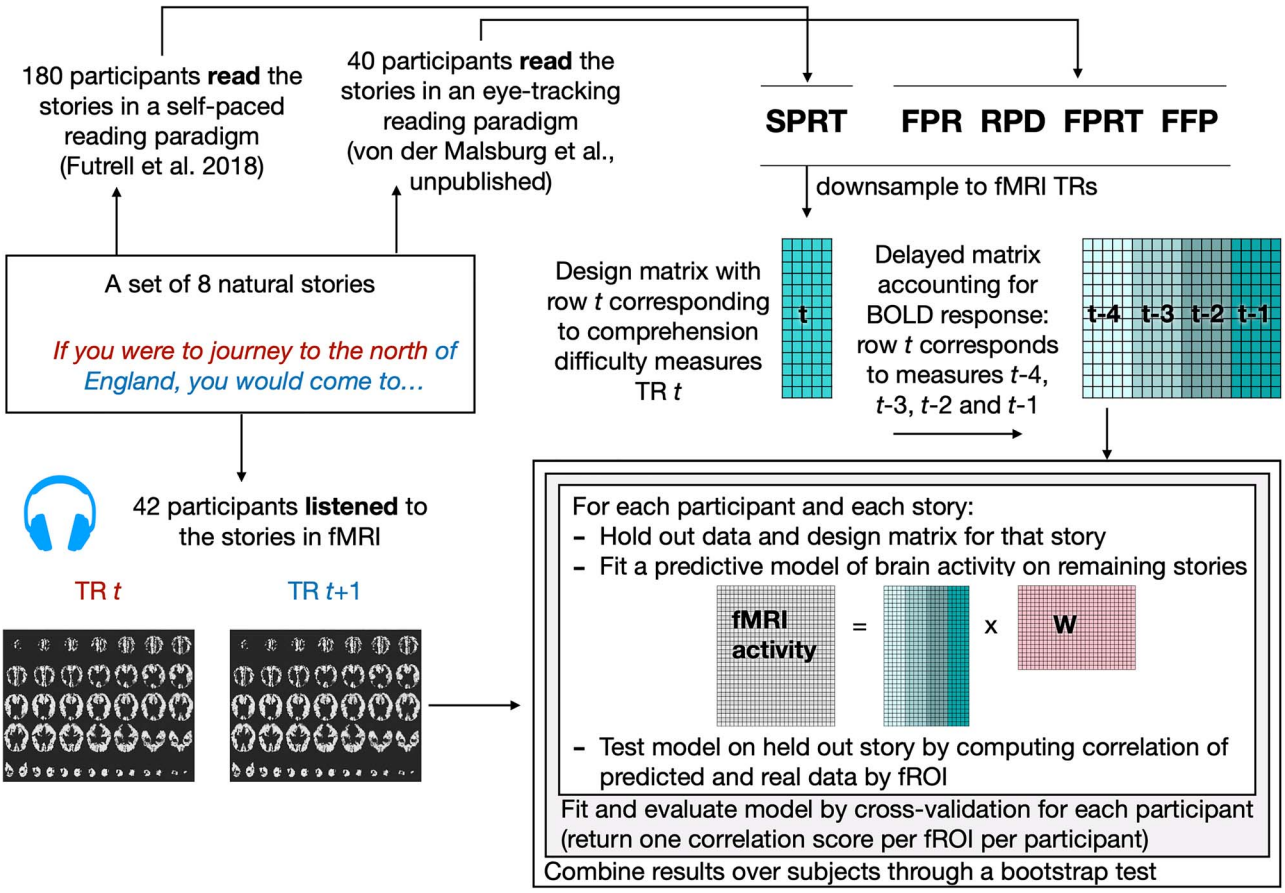

Figure 2. Diagram of approach detailing the combination of data from 3 experiments (fMRI, SPR, and ET) with an encoding model approach. Comprehension difficulty measures are subsampled to the timing of the fMRI TRs and delayed to account for the BOLD response. For each subject, a cross-validation procedure is used where a story is held-out and a predictive model of brain activity as a linear combination of the comprehension difficulty measures is learned. The model is tested on the held-out story. Correlation of predicted and real data is computed for the held-out story; these values are then averaged across all cross-validation folds, resulting in an average correlation by subject and fROI (as well as fROI group). The cross-subjects results are combined using a bootstrap test.

number of stories heard by each participant. In particular, the following procedure was repeated 50 000 times: a set of N ($N = 35$ or $N = 42$ depending on the MD localizer used, as detailed in Methods) participants was sampled with replacement from the original N participants, using a probability distribution where the probability of selecting a subject is proportional to the number of stories they heard. Correlation values of the sampled N participants were averaged for each fROI. 90%, 95%, 98%, and 99% confidence intervals were constructed from the 50 000 samples. Finally, for each fROI, a P-value was obtained from the bootstrap empirical distribution by computing the proportion of samples that were smaller than 0. The Holm–Bonferroni method was applied to correct for multiple hypothesis testing (Holm 1979). In this context where we have a relatively low number of hypotheses (one for every fROI and for every fROI group), controlling for the family-wise error rate (e.g., by using Holm–Bonferroni, as we do here) is more appropriate than controlling the false discovery rate. Given that normalized correlations are a rescaling of the unnormalized correlations, we apply a single test for the unnormalized correlations (with the results of the normalized correlations being the same).

Comparing the Two Networks

To evaluate whether the average brain-behavior correlation across participants is higher in the Language network than in the MD network we again ran a bootstrap test. First, we computed for each participant the difference between the average correlation in the LH Language and MD fROI group. This resulted in N ($N = 35$ or $N = 42$) values. We then repeated the following procedure 50 000 times: a set of N participants was sampled with replacement from the original N participants, and the LH difference for the set of N participants were averaged. This set of 50 000 samples yielded an empirical distribution from which we computed a P-value. We repeated this test to obtain a P value for the difference between the RH Language and MD fROI group. We included these two P-values in the multiple testing correction mentioned in the previous section.

Results

The unnormalized correlations between online comprehension difficulty and BOLD activity in the networks of interest are low, in part because of the low signal-to-noise ratio of fMRI.

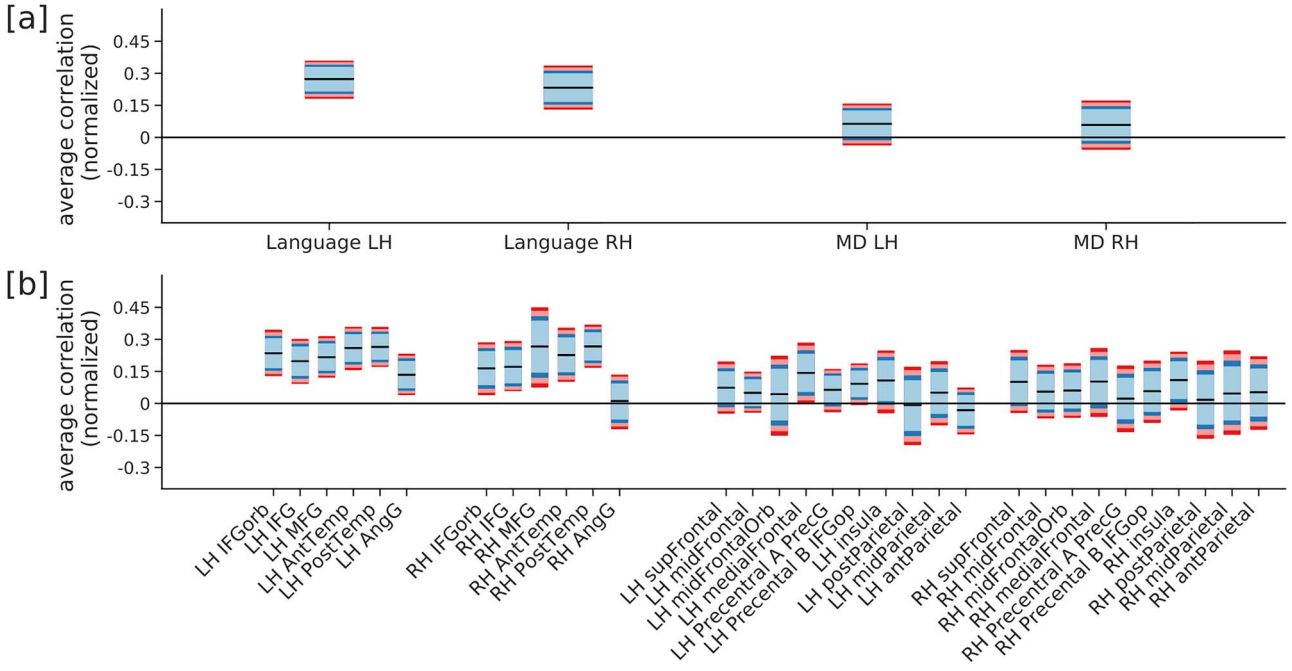

Figure 3. Average normalized correlation between activity predicted as a function of comprehension difficulty (estimated using a combination of SPR times and ET measures) and real held-out activity, for [*a*] each fROI group and [*b*] each fROI. The MD fROIs were localized using the visuo-spatial memory task (available for 35 subjects; see Supplemental for an alternative analysis where the *Nonwords* > *Sentences* contrast, available for all participants, was used as the MD localizer). Performance was averaged across the 35 participants and bootstrap confidence intervals were constructed (Light Blue: 90%, Blue: 95%, Light Red: 98%, Red: 99%). Reading times predict the activity in left and right language fROIs, but not in MD fROIs.

For this reason, we interpreted the size of these correlations by taking into account a metric measure of signal reliability based on inter-subject correlation of the fMRI signals, effectively performing a noise ceiling correction (see Methods; Hsu et al. 2004; Lescroart et al. 2015; Blank and Fedorenko 2017; Lescroart and Gallant 2019). This choice of a ceiling metric leads to conservative normalized correlations: Blank and Fedorenko (2017) show that within-subject correlations are lower than between-subject correlations on this task, and would consequently lead to larger normalized correlations.

Comprehension difficulty measures predicted BOLD activity in the language network, based on either unnormalized or normalized correlations. Average (normalized) prediction performance is significantly greater than chance (family-wise error rate controlled at 0.01) in both the LH and RH language network (Fig. 3*a*), including in most individual fROIs (the bilateral IFGorb, IFG, MFG, AntTemp, and PostTemp fROIs, and the left AngG fROI; Fig. 3*b*). In contrast, comprehension difficulty measures did not significantly predict activity in the MD network, either when averaging across fROIs within the LH or RH, or in any individual MD fROI. (It is worth noting that we chose to run the significance tests on the normalized correlations but running it on the unnormalized correlations, shown in Supplementary Figure 4, would lead to a similar result since the intervals are normalized by a constant, as can be judged by the similarity of the confidence intervals.)

To directly compare between the language and the MD network, we estimated a *P*-value for a 2-sample test by first computing the difference between the prediction performance in the language and MD networks for each subject and then using a bootstrap procedure. We find that the average unnormalized correlation for the language network in each hemisphere is significantly larger than the unnormalized correlation for the MD network ($P = 2 \times 10^{-5}$ for LH and $P = 2 \times 10^{-5}$ for RH). Family-wise error rate was controlled at 0.01.

We additionally performed 2 other versions of this analysis: one, where we included all participants ($N = 42$) and used the *Nonword* > *Sentences* localizer for the MD fROIs (instead of the visuo-spatial *Hard* > *Easy* contrast used the main analysis with $N = 35$), and another where we constrained that same analysis to a subset of those participants ($N = 24$) with near-perfect accuracies on the comprehension questions for the story listening task. The results of both of these additional analyses mirrored the results of the main analysis (with a small variation in the correlation values and the confidence interval leading to a change in significance in the right MFG language fROI in the high-accuracy subjects analysis) (see Table 2 and Supplementary Figs 2 and 3).

Discussion

In this study, we tested whether language comprehension—in addition to language-specific resources—recruits domain-general executive mechanisms. To this end, we examined the relationship between behavioral and neural measures of incremental comprehension difficulty during naturalistic language processing: behavioral comprehension difficulty was estimated with 2 commonly used approaches (SPR times and ET fixation durations); and neural recruitment was quantified from fMRI BOLD activity during auditory language comprehension in domain-general and language-selective functional networks that have been previously implicated in language comprehension. We found that whereas neural activity in the fronto-temporal language-selective network (Fedorenko et al. 2011; Fedorenko and Thompson-Schill 2014;

Table 2 Summary of results across fROI groups (bolded) and fROIs, for the main analysis and the 2 additional supplementary analyses

		Main analysis	Analysis with the *Nonwords* > *Sentences* MD localizer (Supplementary Fig. 2)	Analysis with only high-performing participants (Supplementary Fig. 3)
Network effects	**LH Language**	**X**	**X**	**X**
	RH Language	**X**	**X**	**X**
	LH MD			
	RH MD			
Effects for individual language fROIs	LH IFGorb	X	X	X
	LH IFG	X	X	X
	LH MFG	X	X	X
	LH AntTemp	X	X	X
	LH PostTemp	X	X	X
	LH AngG	X	X	X
	RH IFGorb	X	X	X
	RH IFG	X	X	X
	RH MFG	X	X	
	RH AntTemp	X	X	X
	RH PostTemp	X	X	X
	RH AngG			
Effects for individual MD fROIs	LH supFrontal			
	LH midFrontal			
	LH midFrontalOrb			
	LH medialFrontal			
	LH Precentral A PrecG			
	LH Precental B IFGop			
	LH insula			
	LH postParietal			
	LH midParietal			
	LH antParietal			
	RH supFrontal			
	RH midFrontal			
	RH midFrontalOrb			
	RH medialFrontal			
	RH Precentral A PrecG			
	RH Precental B IFGop			
	RH insula			
	RH postParietal			
	RH midParietal			
	RH antParietal			

Note: X marks a significantly higher than 0 correlation. The 3 sets of results are extremely similar, except for variation in only one fROI.

Braga et al. 2020) was predicted by behavioral measures of incremental comprehension difficulty, activity in the domain-general MD network (Duncan 2010; Assem et al. 2020b) was not predicted by these measures. Furthermore, the difference between the language and MD networks was reliable: the mean prediction performance was significantly higher in the LH or RH language network than in the LH and RH MD network, respectively. (Note that, although the prediction performance was not significantly higher than chance in any of the 20 MD fROIs, the average correlation across participants was still positive in 18 of those fROIs. Had those fROIs been independent, 18 out of 20 fROIs showing a small positive value would have been surprising and would have suggested that with more power, those effects could become significant. However, given that the MD network is strongly functionally integrated with high correlations among the regions in BOLD signal fluctuation patterns during naturalistic cognition (Blank et al. 2014; Paunov et al. 2019) and in effect sizes during task paradigms (Mineroff et al. 2018; Assem et al. 2020a) a possible explanation for the nonsignificant positive prediction in those fROIs is the presence of shared noise.) The lack of a reliable correlation between behavioral comprehension difficulty and neural activity in the MD network conflicts with previous studies reporting (putative) MD activity (most past studies did not include an independent localizer for the MD network as needed to interpret the observed effects without relying on reverse inference reasoning; Poldrack 2006) during some language tasks and the sensitivity of the MD regions to linguistic manipulations of processing difficulty (e.g., January et al. 2009; Kuperberg et al. 2003; McMillan et al. 2013; Nieuwland et al. 2012; Novais-Santos et al. 2007; Peelle et al. 2010; Rodd et al. 2005).

Our investigation complements the Henderson et al. (2015) study discussed in the introduction, which related behavioral (ET) and neural measures of language comprehension obtained

from the same participants. Similar to our study, Henderson et al. (2015) observed effects in brain areas in left temporal lobe commonly associated with language comprehension, that is, putative parts of the "core language network," and they did not observe effects in the frontal or parietal MD regions. However, they relied on a traditional group-based analytic approach, which relies on voxel-wise correspondence across individuals and does not take into account the well-known inter-individual variation in the precise locations of functional areas in the association cortex (e.g., Fischl et al. 2008; Frost and Goebel 2012; Tahmasebi et al. 2012; Vázquez-Rodríguez et al. 2019). Because of the resulting low sensitivity of such analyses (e.g., Nieto-Castañón and Fedorenko 2012), Henderson et al. (2015) may have missed the effects within the MD network. In the current study, to maximize the probability of detecting a relationship between behavioral measures and MD activity, we functionally localized MD areas (as well as core language areas) in each individual participant (Fedorenko et al. 2010). This strategy reduces the risk of obtaining a false negative for the MD network.

Our use of separate participant pools in the behavioral studies vs. the fMRI study further eliminates many nonlinguistic confounds (such as attention and motor control processes related to eye-movements or button presses) that are difficult to avoid when the behavioral and neural measures come from the same individuals as in Henderson et al. (2015). The observed relationship between reading latencies and brain activity is thus most plausibly due to comprehension demands related to the "linguistic properties" of the story stimuli, including lexical-level variables (e.g., word length or frequency; Howes and Solomon 1951; Hudson and Bergman 1985; see Futrell et al. 2018, for evidence that these effects are captured by the RTs in the SPR measure used here) and variables that affect the ease of establishing syntactic and semantic dependencies among words (e.g., dependency length, structural frequency, or the presence of syntactic ambiguity; Dopkins et al. 1992; Sturt 2007; Smith and Levy 2013). Reading latencies also reflect linguistic properties that affect discourse-level processing (Gordon and Chan 1995; Gordon and Scearce 1995; Chambers and Smyth 1998), although prior evidence has suggested that the core language network is insensitive to discourse structure (e.g., Ferstl and Von Cramon 2001; Lerner et al. 2011; Blank and Fedorenko 2020; Jacoby and Fedorenko 2020).

Although we consider our use of separate cohorts of participants and different language comprehension modalities (reading for behavioral measures, and listening for the fMRI measure) to be a strength (allowing us to isolate language comprehension difficulty from perceptual-level and modality-specific effects), we concede that this approach may have limitations. For example, aspects of comprehension that are specific to listening (e.g., the increased difficulty in following a speaker who is too fast or too slow, or prosodic effects) are not included in our reading measures of comprehension difficulty and deserve further investigation. As emphasized earlier, the current study focused on comprehension demands related directly to the "properties of linguistic materials," which are similar between the 2 presentation modalities. Another potential limitation of the design where the behavioral measure comes from an independent group of participants is reduction in statistical power. However, this limitation is mitigated by the fact that we observed large and robust predictive effects in the language network.

The current findings support the hypotheses that 1) behavioral responses to language stimuli reflect computational load in language comprehension mechanisms (Just and Carpenter 1980; Rayner 1998; Reichle et al. 1998); 2) language comprehension difficulty generalizes across participants, task demands, and modality of presentation (Demberg and Keller 2008; Frank and Bod 2011; Vagharchakian et al. 2012; Smith and Levy 2013; Fedorenko et al. 2016; Scott et al. 2017; Deniz et al. 2019; Shain 2019; Shain and Schuler 2019; Staub and Goddard 2019; Staub 2020); and 3) processing mechanisms that give rise to measurable reading delays reside in the language-selective cortical network, rather than in the domain-general executive control network. In this way, our results echo and reinforce those of Henderson et al. (2015): we replicate their finding of a relationship between reading latencies and neural activity in the left-hemisphere middle and superior temporal lobe, and extend it to other parts of the language network, including the language-responsive areas in the left inferior frontal cortex and, to a lesser extent, the left angular gyrus, as well as the right-hemisphere homotopic language areas. These more widespread effects across the language network are plausibly due to the increased sensitivity gained from participant-specific functional localization (e.g., Nieto-Castañón and Fedorenko 2012).

At present, much behavioral language research is disconnected from cognitive neuroscience efforts to understand the architecture of language comprehension, despite 1) the fact that these 2 enterprises share the same goal—to understand the computations that support language comprehension, and 2) the fact that a link between behavioral measures of language comprehension—or the mental states they correspond to—and neural correlates of language comprehension is a fundamental assumption of psycholinguistics (e.g., Just and Carpenter 1980). Indeed, except for Henderson et al. (2015) and the current paper, cognitive neuroscientists have not typically used direct and continuous behavioral measures to model brain activity during language comprehension (see e.g., Supplementary Table 1 for fMRI studies that have used naturalistic linguistic materials and which have typically used linguistic features as predictors of neural activity, often without first establishing a link between those features and *behavioral* measures). The current paper connects the psycholinguistic and cognitive neuroscience literatures, and in so doing contributes to both fields. For psycholinguistics, our results validate widely used behavioral measures as indeed revealing the underlying activity of language(-selective) comprehension mechanisms. For cognitive neuroscience, our results indicate that, even using a broad (all-encompassing) and theory-neutral estimate of comprehension difficulty, language processing recruits primarily cortical circuits that specialize for this purpose, and that domain-general executive mechanisms are generally not recruited during naturalistic sentence comprehension.

This work thus sheds new light on the role of the domain-general MD network in language comprehension, and on the division of labor between these domain-general mechanisms and the language-selective ones. In particular, regions of the MD network have been shown to be sensitive to linguistic difficulty across diverse manipulations (see Fedorenko 2014, for a review). However, almost all prior evidence has come from traditional, task-based experimental paradigms that present participants with linguistic manipulations that do not commonly occur in real-life comprehension scenarios (like ambiguous words that are not disambiguated by the context, or nonlocal dependencies; e.g., Rodd et al. 2005; Novais-Santos et al. 2007; January et al. 2009; Peelle et al. 2010) and ask them to solve "artificial" tasks, such as making similarity judgments or deciding whether a sentence matches a picture. Although the stories

used in the current study were modified to include challenging linguistic phenomena in order to increase variability in processing demands and increase the chances of engaging executive resources, the only "task" required of participants during story listening was comprehension of the narratives. The fact that we do not find a relationship between comprehension difficulty and the MD network's activity in our study suggests that the MD network's contribution to language comprehension may be restricted to artificial scenarios, where language is effectively turned into problem solving (Wright et al. 2011; Diachek et al. 2020). In line with this conjecture, Blank and Fedorenko (2017) showed that the activity in MD regions is not strongly correlated across participants or reliable within participants during comprehension, indicating that the MD system is not closely tracking the language stimulus, and Shain et al. (2020) showed that activity in the MD regions during comprehension does not correlate with the psycholinguistic construct of "surprisal," the moment-by-moment unpredictability of linguistic input (see also Blanco-Elorrieta and Pylkkänen 2017, for evidence of less MD engagement during a more naturalistic production paradigm). Whereas the MD network may play some role during language processing (perhaps modulating overall alertness or attention) our results as well as others mentioned above suggest that this system is not directly involved in linguistic computations related to lexical access or constructing inter-word syntactic and semantic dependencies.

Our results also suggest that similarities between language processing and other kinds of processing in other domains (e.g., theoretical constructs in mathematics, music, and computer programming resembling those in natural language syntax) do not entail shared neural circuitry (see also Fedorenko and Blank 2020, for a recent discussion). In particular, the fact that multiple domains require hierarchical combinatorial processing does not mean that the same circuits are engaged across those domains. Rather, constructing hierarchical structures, predictive coding, working memory storage and retrieval of information, and other processes that may be necessary in multiple domains of cognition appear to be implemented within domain-specialized systems, like the core language network that supports language processing.

In conclusion, we found that whereas neural activity in the fronto-temporal language network is predicted by behavioral signatures of incremental comprehension difficulty, activity in the domain-general fronto-parietal MD network is not.

Supplementary Material

Supplementary material can be found at *Cerebral Cortex* online.

Notes

We acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research, MIT. For technical support during scanning, the authors thank Steve Shannon, Christina Triantafyllou, and Atsushi Takahashi. We thank Anastasia Vishnevetsky and Steve Piantadosi for help in constructing the story materials, Nancy Kanwisher for help in recording the story materials, Kris Fedorenko for editing the recordings, Zach Mineroff for help in collecting the fMRI data and preparing the data for analyses, Hal Tily for help in collecting the self-paced reading data, Kyle Mahowald for help in preprocessing the self-paced reading data, Wade Shen and Jeanne Gallée for help in aligning the story transcripts to the auditory files, Jessica Chen and Hannah Small for help with visualizing ROI projections. *Conflicts of Interest:* No financial interests or conf licts of interest.

Authors' Contributions

LW, IB, EG, and EF designed research. All authors performed research: LW, IB, and EF collected, preprocessed, and analyzed fMRI data; RF, IB, and EG collected, preprocessed, and analyzed the self-paced reading data; RL, TM, and NS collected, preprocessed, and analyzed the eye-tracking data. All authors interpreted the data. LW and EF wrote the manuscript. IB, CS, RF, RL, TM, and EG provided comments.

Funding

L.W. was supported by startup funds at Carnegie Mellon University. R.F. was supported by NSF BCS DDRI grant number 1551543. T.M. was supported by a Feodor Lynen Research Fellowship awarded to him by the Alexander von Humboldt Foundation and by NIH NICHD grant HD065829 awarded to R.L. and Keith Rayner. R.L. was also supported by an Alfred P. Sloan Research Fellowship and a Paul and Lilah Newton Brain Science Award. E.F. was supported by NIH grants HD057522 NICHD, DC016607 NIDCD, and DC016950 NIDCD, a grant from the Simons Foundation to the Simons Center for the Social Brain at MIT, and by funds from the Department of Brain and Cognitive Sciences and the McGovern Institute for Brain Research at MIT.

References

Aaronson D, Scarborough HS. 1977. Performance theories for sentence coding: some quantitative models. *J Verbal Learning Verbal Behav*. 16(3):277–303.

Abney SP, Johnson M. 1991. Memory requirements and local ambiguities of parsing strategies. *J Psycholinguist Res*. 20(3):233–250.

Anderson ML. 2010. Neural reuse: a fundamental organizational principle of the brain. *Behav Brain Sci*. 33(4):245–266.

Assem M, Blank IA, Mineroff Z, Ademoğlu A, Fedorenko E. 2020a. Activity in the fronto-parietal multiple-demand network is robustly associated with individual differences in working memory and fluid intelligence. *Cortex*.

Assem M, Glasser MF, Van Essen DC, Duncan J. 2020b. A domain-general cognitive core defined in multimodally Parcellated human cortex. *Cereb Cortex*.

Ayyash D, Malik Moraleda SM, Galleé JMZ, Jouravlev O, Fedorenko E. n.d.The universal language network: a cross-linguistic investigation spanning 41 languages and 10 language families. In Preparation.

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

Bhattasali S, Hale J, Pallier C, Brennan JR, Luh W-M, Spreng RN. 2018. Differentiating phrase structure parsing and memory retrieval in the brain. *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, Salt Lake City, Utah, January 4–7, 2018, 2012, 74–80.

Bilenko NY, Grindrod CM, Myers EB, Blumstein SE. 2008. Neural correlates of semantic competition during processing of ambiguous words. *J Cogn Neurosci*. 21(5):960–975.

Binder JR. 1997. Neuroanatomy of language processing studied with functional MRI. *Clin Neurosci*. 4(2):87–94.

Blanco-Elorrieta E, Pylkkänen L. 2017. Bilingual language switching in the laboratory versus in the wild: the spatiotemporal dynamics of adaptive language control. *J Neurosci*. 37(37):9022–9036.

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

Blank I, Fedorenko E. 2017. Domain-general brain regions do not track linguistic input as closely as language-selective regions. *J Neurosci*. 3642–3616.

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

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

Blumstein SE, Amso D. 2013. Dynamic functional organization of language: insights from functional neuroimaging. *Perspect Psychol Sci*. 8(1):44–48.

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

Braze D, Mencl WE, Tabor W, Pugh KR, Constable RT, Fulbright RK, Magnuson JS, Van Dyke JA, Shankweiler DP. 2011. Unification of sentence processing via ear and eye: an fMRI study. *Cortex*. 47(4):416–431.

Brennan J, Nir Y, Hasson U, Malach R, Heeger DJ, Pylkkänen L, Manuscript A, Structures T. 2010. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain Lang*. 6(2):247–253.

Brennan J, Nir Y, Hasson U, Malach R, Heeger DJ, Pylkkänen L. 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain Lang*. 120(2):163–173.

Brennan J, Pylkkänen L. 2012. The time-course and spatial distribution of brain activity associated with sentence processing. *Neuroimage*. 60(2):1139–1148.

Brennan JR, Stabler EP, Van Wagenen SE, Luh WM, Hale JT. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain Lang*. 157–158:81–94.

Broca PP. 1861. Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). In: *Bull Soc Anat*.

Campbell KL, Tyler LK. 2018. Language-related domain-specific and domain-general systems in the human brain. *Curr Opin Behav Sci*. 21:132–137.

Chambers CG, Smyth R. 1998. Structural parallelism and discourse coherence: a test of centering theory. *J Mem Lang*. 39(4):593–608.

Clifton C, Frazier L. 1989. In: Carlson GN, Tanenhaus MK, editors. *Comprehending sentences with long-distance dependencies BT – linguistic structure in language processing*. Springer Netherlands, pp. 273–317.

Cohen AL. 2013. Software for the automatic correction of recorded eye fixation locations in reading experiments. *Behav Res Methods*. 45(3):679–683.

Cohen J, Cohen P, West SG, Aiken LS. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

D'Esposito M, Postle BR. 2015. The cognitive neuroscience of working memory. *Annu Rev Psychol*. 66(1):115–142.

De Deyne S, Navarro DJ, Storms G. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behav Res Methods*. 42(2):480–498.

de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE. 2017. The hierarchical cortical Organization of Human Speech Processing. *J Neurosci*. 37(27):6539–6557.

Dehghani M, Boghrati R, Man K, Hoover J, Gimbel SI, Vaswani A, Zevin JD, Immordino-Yang MH, Gordon AS, Damasio A *et al*. 2017. Decoding the neural representation of story meanings across languages. *Hum Brain Mapp*. 38(12):6096–6106.

Demberg V, Keller F. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*. 109(2):193–210.

Deniz F, Nunez-Elizalde AO, Huth AG, Gallant JL. 2019. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *J Neurosci*. 39(39):7722–7736.

Desai RH, Choi W, Lai VT, Henderson JM. 2016. Toward semantics in the wild: activation to manipulable nouns in naturalistic reading. *J Neurosci*. 36(14):4050–4055.

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. *J Neurosci*. 40(23):4536–4550.

Dopkins S, Morris RK, Rayner K. 1992. Lexical ambiguity and eye fixations in reading: a test of competing models of lexical ambiguity resolution. *J Mem Lang*. 31(4):461–476.

Dronkers NF, Wilkins DP, Van Valin RD, Redfern BB, Jaeger JJ. 2004. Lesion analysis of the brain areas involved in language comprehension. *Cognition*. 92(1–2):145–177.

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

Duncan J. 2013. The structure of cognition: attentional episodes in mind and brain. *Neuron*. 80(1):132–137.

Duncan J, Assem M, Shashidhara S. 2020. Integrated intelligence from distributed brain activity. *Trends Cogn Sci*. 24(10):838–852.

Duncan J, Owen AM. 2000. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends Neurosci*.

Eklund A, Andersson M, Josephson C, Johannesson M, Knutsson H. 2012. Does parametric fMRI analysis with SPM yield valid results?-an empirical study of 1484 rest datasets. *Neuroimage*. 61(3):565–578.

Fadiga L, Craighero L, D'Ausilio A. 2009. Broca's area in language, action, and music. *Ann N Y Acad Sci*. 1169(1):448–458.

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

Fedorenko E, Hsieh P-J, Nieto-Castanon A, Whitfield-Gabrieli S, Kanwisher N. 2010. New method for f{MRI} investigations of language: defining {ROI}s functionally in individual subjects. *J Neurophysiol*. 104(2):1177–1194.

Fedorenko E, Nieto-Castanon A, Kanwisher N. 2012. Lexical and syntactic representations in the brain: an f{MRI} investigation with multi-voxel pattern analyses. *Neuropsychologia*. 50(4):499–513.

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

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

Fedorenko E, Blank IA. 2020. Broca's area is not a natural kind. *Trends Cogn Sci*. 24(4):270–284.
Fedorenko E, Duncan J, Kanwisher N. 2013. Broad domain generality in focal regions of frontal and parietal cortex. *Proc Natl Acad Sci*. 110(41):16616–16621.
Fedorenko E, Gibson E, Rohde D. 2006. The nature of working memory capacity in sentence comprehension: evidence against domain-specific working memory resources. *J Mem Lang*. 54(4):541–553.
Fedorenko E, Gibson E, Rohde D. 2007. The nature of working memory in linguistic, arithmetic and spatial integration processes. *J Mem Lang*. 56(2):246–265.
Fedorenko E, Blank I, Siegelman M, Mineroff Z. 2020b. Word meanings and sentence structure recruit the same set of fronto-temporal regions during comprehension. *BioRxiv*. 477851.
Fedorenko E, Scott TL, Brunner P, Coon WG, Pritchett B, Schalk G, Kanwisher N. 2016. Neural correlate of the construction of sentence meaning. *Proc Natl Acad Sci*. 113(41): E6256–E6262.
Fedorenko E, Thompson-Schill SL. 2014. Reworking the language network. *Trends Cogn Sci*. 18(3):120–127.
Fedorenko E, Varley R. 2016. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Ann N Y Acad Sci*. 1369(1):132–153.
Ferstl EC, Von Cramon DY. 2001. The role of coherence and cohesion in text comprehension: an event-related fMRI study. *Cogn Brain Res*. 11(3):325–340.
Fiebach CJ, Vos SH, Friederici AD. 2004. Neural correlates of syntactic ambiguity in sentence comprehension for low and high span readers. *J Cogn Neurosci*. 16(9):1562–1575.
Fischl B, Rajendran N, Busa E, Augustinack J, Hinds O, Yeo BTT, Mohlberg H, Amunts K, Zilles K. 2008. Cortical folding patterns and predicting cytoarchitecture. *Cereb Cortex*. 18(8):1973–1980.
Fitch WT, Martins MD. 2014. Hierarchical processing in music, language, and action: Lashley revisited. *Ann N Y Acad Sci*. 1316(1):87–104.
Frank SL, Bod R. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychol Sci*. 22(6):829–834.
Frazier L, Rayner K. 1987. Resolution of syntactic category ambiguities: eye movements in parsing lexically ambiguous sentences. *J Mem Lang*. 26(5):505–526.
Friederici AD, Rüschemeyer S-A, Hahne A, Fiebach CJ. 2003. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. *Cerebral Cortex (New York, NY : 1991)*. 13(2):170–177.
Friedrich R, Friederici AD. 2009. Mathematical logic in the human brain: syntax. *PLoS One*. 4(5):e5599.
Frost MA, Goebel R. 2012. Measuring structural-functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. *Neuroimage*. 59(2):1369–1381.
Futrell R, Gibson E, Tily HJ, Blank I, Vishnevetsky A, Piantadosi ST, Fedorenko E. 2020. The natural stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Lang Resour Eval*. 1–5.
Futrell R, Gibson E, Tily HJH, Blank I, Vishnevetsky A, Piantadosi ST, Fedorenko E. 2018. The natural stories corpus. *Proceedings of the 11th Language Resources and Evaluation Conference*, 1–16.
Gernsbacher MA. 1993. Less skilled readers have less efficient suppression mechanisms. *Psychol Sci*. 49(5):294–298.
Geschwind N. 1970. The organization of language and the brain. *Science*. 170(3961):940–944.
Gibson E. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*. 68(1):1–76.
Gibson E. 2000. The dependency locality theory: a distance-based theory of linguistic complexity. *Image, Lang, Brain*. 95–126.
Gibson EAF. 1991. *A computational theory of human linguistic processing: memory limitations and processsing breakdown*, p. 206.
Gordon PC, Chan D. 1995. Pronouns, passives, and discourse coherence. *J Mem Lang*. 34(2):216–231.
Gordon PC, Hendrick R, Levine WH. 2002. Memory-load interference in syntactic processing. *Psychol Sci*. 13(5):425–430.
Gordon PC, Scearce KA. 1995. Pronominalization and discourse coherence, discourse structure and pronoun interpretation. *Mem Cognit*. 23(3):313–323.
Grodner D, Gibson E, Tunstall S. 2002. Syntactic complexity in ambiguity resolution. *J Mem Lang*. 46(2):267–295.
Hagoort P. 2019. The neurobiology of language beyond single-word processing. *Science*. 366(6461).
Hale JT, Lutz DE, Luh W, Brennan JR, Arbor A. 2015. Modeling fMRI time courses with linguistic structure at various grain sizes. *Proceedings of CMCL*, 89–97.
Handwerker DA, Ollinger JM, D'Esposito M. 2004. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*. 21(4):1639–1651.
Hasson U, Egidi G, Marelli M, Willems RM. 2018. Grounding the neurobiology of language in first principles: the necessity of non-language-centric explanations for language comprehension. *Cognition*. 180:135–157.
Hasson U, Honey CJ. 2012. Future trends in {N}euroimaging: {N}eural processes as expressed within real-life contexts. *Neuroimage*. 62(2):1272–1278.
Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N. 2008. A hierarchy of temporal receptive windows in human cortex. *J Neurosci*. 28(10):2539–2550.
Heim S, Eickhoff SB, Amunts K. 2008. Specialisation in Broca's region for semantic, phonological, and syntactic fluency? *Neuroimage*. 40(3):1362–1368.
Henderson JM, Choi W, Lowder MW, Ferreira F. 2016. Language structure in the brain: a fixation-related fMRI study of syntactic surprisal in reading. *Neuroimage*. 132:293–300.
Henderson JM, Choi W, Luke SG, Desai RH. 2015. Neural correlates of fixation duration in natural reading: evidence from fixation-related fMRI. *Neuroimage*. 119:390–397.
Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scand J Stat*.
Howes DH, Solomon RL. 1951. Visual duration threshold as a function of word-probability. *J Exp Psychol*.
Hsu A, Borst A, Theunissen F. 2004. Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Comput Neural Syst*. 15(2): 91–109.
Hsu NS, Novick JM. 2016. Dynamic engagement of cognitive control modulates recovery from misinterpretation during real-time language processing. *Psychol Sci*. 27(4):572–582.
Hudson PTW, Bergman MW. 1985. Lexical knowledge in word recognition: word length and word frequency in naming and lexical decision tasks. *J Mem Lang*. 24(1):46–58.
Hugdahl K, Raichle ME, Mitra A, Specht K. 2015. On the existence of a generalized non-specific task-dependent network. *Front Hum Neurosci*. 1–5.

Humphries C, Binder JR, Medler DA, Liebenthal E. 2006. Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *J Cogn Neurosci*. 18(4):665–679.

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL, Heer WAD, Griffiths TL, Gallant JL. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*. 532(7600):453–458.

Ivanova A, Siegelman M, Cheung C, Pongos A, Kean H, Fedorenko E. n.d. Effect of task on sentence processing. In Preparation.

Jacoby N, Fedorenko E. 2020. Discourse-level comprehension engages medial frontal theory of mind brain regions even for expository texts. *Lang, Cognit Neurosci*. 35(6):780–796.

January D, Trueswell JC, Thompson-Schill SL. 2009. Co-localization of stroop and syntactic ambiguity resolution in Broca's area: implications for the neural basis of sentence processing. *J Cogn Neurosci*. 21(12):2434–2444.

Johnson-Laird PN. 1983. *Mental models: towards a cognitive science of language, inference, and consciousness* (Issue 6). Harvard University Press.

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

Jung-Beeman M. 2005. Bilateral brain processes for comprehending natural language. *Trends Cogn Sci*. 9(11):512–518.

Just MA, Carpenter PA, Woolley JD. 1982. Paradigms and processes in reading comprehension. J Exp Psychol. *General*. 111(2):228–238.

Just MA, Carpenter PA. 1980. A theory of reading: from eye fixations to comprehension. *Psychol Rev*. 87(4):329–354.

Kaakinen JK, Hyönä J. 2010. Task effects on eye movements during reading. *J Exp Psychol Learn Mem Cogn*. 36(6):1561–1566.

Kaan E, Swaab TY. 2002. The brain circuitry of syntactic comprehension. *Trends Cogn Sci*. 6(8):350–356.

Keller TA, Carpenter PA, Just MA. 2001. *The neural bases of sentence comprehension: a fMRI examination of syntactic and lexical processing*, pp. 223–237.

Kennedy A. 2000. Parafoveal processing in word recognition. Q *J Exp Psychol Sect A*. 53(2):429–455.

King J, Just MA. 1991. Individual differences in syntactic processing: the role of working memory. *J Mem Lang*. 30(5):580–602.

Klein R, Farrell M. 1989. Search performance without eye movements. *Percept Psychophys*. 46(5):476–482.

Kuperberg GR, Holcomb PJ, Sitnikova T, Greve D, Dale AM, Caplan D. 2003. Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *J Cogn Neurosci*. 15(2):272–293.

Kuperberg GR, Sitnikova T, Lakshmanan BM. 2008. Neuroanatomical distinctions within the semantic system during sentence comprehension: evidence from functional magnetic resonance imaging. *Neuroimage*. 40(1):367–388.

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

Lescroart MD, Gallant JL. 2019. Human scene-selective areas represent 3D configurations of surfaces. *Neuron*. 101(1):178–192-e7.

Lescroart MD, Stansbury DE, Gallant JL. 2015. Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. *Front Comput Neurosci*. 9:135.

Levy R. 2008. Expectation-based syntactic comprehension. *Cognition*. 106(3):1126–1177.

Lewis RL, Vasishth S. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognit Sci*. 29(3):375–419.

Lewis RL, Vasishth S, Van Dyke JA. 2006. Computational principles of working memory in sentence comprehension. *Trends Cogn Sci*. 10(10):447–454.

Lopopolo A, Frank SL, Van Den Bosch A, Willems RM. 2017. Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLoS One*. 12(5):1–18.

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.

Mazoyer BM, Tzourio N, Frak V, Syrota A, Murayama N, Levrier O, Salamon G, Dehaene S, Cohen L, Mehler J. 1993. The cortical representation of speech. *J Cogn Neurosci*. 5(4):467–479.

McElree B. 2000. Sentence comprehension is mediated by content-addressable memory structures. *J Psycholinguist Res*. 29(2):111–123.

McElree B. 2001. Working memory and focal attention. *J Exp Psychol Learn Mem Cogn*.

McMillan CT, Clark R, Gunawardena D, Ryant N, Grossman M. 2012. FMRI evidence for strategic decision-making during resolution of pronoun reference. *Neuropsychologia*. 50(5): 674–687.

McMillan CT, Coleman D, Clark R, Liang TW, Gross RG, Grossman M. 2013. Converging evidence for the processing costs associated with ambiguous quantifier comprehension. *Front Psychol*. 4(APR):1–10.

Menke J, Martinez TR. 2004. Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, 2, 1331–1335.

Mesulam MM. 1998. From sensation to cognition. *Brain*. 121(6):1013–1052.

Miller EK, Cohen JD. 2001. An integrative theory of prefrontal cortex function. *Annu Rev Neurosci*. 24(1):167–202.

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

Mitchell DC. 1984. An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In: *New Methods in Reading Comprehension Research*, pp. 69–89.

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

Monti MM, Parsons LM, Osherson DN. 2009. The boundaries of language and thought in deductive inference. *Proc Natl Acad Sci U S A*. 106(30):12554–12559.

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

Murphy B, Hale J, Brennan J. 2016. Grammatical Relations in the Listening Brain. *Poster at PRNI 2016, Pattern Recognition and Neuroimaging Conference*.

Naselaris T, Kay KN, Nishimoto S, Gallant JL. 2011. Encoding and decoding in fMRI. *Neuroimage*. 56(2):400–410.

Nelson MJ, El Karoui I, Giber K, Yang X, Cohen L, Koopman H, Cash SS, Naccache L, Hale JT, Pallier C *et al*. 2017.

Neurophysiological dynamics of phrase-structure building during sentence processing. *Proc Natl Acad Sci U S A*. 114(18):E3669–E3678.
Nieto-Castañón A, Fedorenko E. 2012. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *Neuroimage*. 63(3):1646–1669.
Nieuwland MS, Martin AE, Carreiras M. 2012. Brain regions that process case: evidence from basque. *Hum Brain Mapp*. 33(11):2509–2520.
Novais-Santos S, Gee J, Shah M, Troiani V, Work M, Grossman M. 2007. Resolving sentence ambiguity with planning and working memory resources: evidence from fMRI. *Neuroimage*. 37(1):361–378.
Novick JM, Kan IP, Trueswell JC, Thompson-Schill SL. 2009. A case for conflict across multiple domains: memory and language impairments following damage to ventrolateral prefrontal cortex. *Cogn Neuropsychol*. 26(6):527–567.
Novick JM, Trueswell JC, Thompson-Schill SL. 2005. Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cogn Affect Behav Neurosci*. 5(3):263–281.
Pallier C, Devauchelle AD, Dehaene S. 2011. Cortical representation of the constituent structure of sentences. *Proc Natl Acad Sci*. 108(6):2522–2527.
Patel AD. 2003. Language, music, syntax and the brain. *Nat Neurosci*. 6(7):674–681.
Patel AD. 2012. Music, language, and the brain. *Music, Lang Brain*.
Paunov AM, Blank IA, Fedorenko E. 2019. Functionally distinct language and theory of mind networks are synchronized at rest and during language comprehension. *J Neurophysiol*. 121(4):1244–1265.
Peelle JE, Troiani V, Wingfield A, Grossman M. 2010. Neural processing during older adults' comprehension of spoken sentences: age differences in resource allocation and connectivity. *Cereb Cortex*. 20(4):773–782.
Pickering MJ, Frisson S. 2001. Processing ambiguous verbs: evidence from eye movements. *J Exp Psychol Learn Mem Cogn*. 27(2):556–573.
Poldrack RA. 2006. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci*. 10(2):59–63.
Posner MI. 1980. Orienting of attention. *Q J Exp Psychol*. 32(1): 3–25.
Posner MI. 2016. Orienting of attention: then and now. *Q J Exp Psychol*. 69(10):1864–1875.
Pritchett BL, Hoeflin C, Koldewyn K, Dechter E, Fedorenko E. 2018. High-level language processing regions are not engaged in action observation or imitation. *J Neurophysiol*. 120(5):2555–2570.
Rasmussen NE, Schuler W. 2018. Left-corner parsing with distributed associative memory produces Surprisal and locality effects. *Cognit Sci*. 42:1009–1042.
Rayner K. 1977. Visual attention in reading: eye movements reflect cognitive processes. *Mem Cognit*. 5(4):443–448.
Rayner K. 1978. Eye movements in reading and information processing. *Psychol Bull*. 85(3):618–660.
Rayner K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychol Bull*. 124(3): 372.
Regev M, Honey CJ, Simony E, Hasson U. 2013. Selective and invariant neural responses to spoken and written narratives. *J Neurosci*. 33(40):15978–15988.
Reichle ED, Pollatsek A, Fisher DL, Rayner K. 1998. Toward a model of eye movement control in reading. *Psychol Rev*.
Remington RW. 1980. Attention and saccadic eye movements. *J Exp Psychol Hum Percept Perform*. 6(4):726–744.
Resnik P. 1992. *Left-corner parsing and psychological plausibility*. COLING, 191.
Rodd JM, Davis MH, Johnsrude IS. 2005. The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cereb Cortex*. 15(8):1261–1269.
Rodd JM, Johnsrude IS, Davis MH. 2010. The role of domain-general frontal systems in language comprehension: evidence from dual-task interference and semantic ambiguity. *Brain Lang*. 115(3):182–188.
Rodriguez A, Granger R. 2016. The grammar of mammalian brain capacity. *Theor Comput Sci*. 633:100–110.
Rogalsky C, Hickok G. 2011. The role of Broca's area in sentence comprehension. *J Cogn Neurosci*. 23(7):1664–1680.
Saxe R, Brett M, Kanwisher N. 2006. Divide and conquer: a defense of functional localizers. *Neuroimage*. 30(4): 1088–1096.
Schotter ER, Tran R, Rayner K. 2014. Don't believe what you read (only once): comprehension is supported by regressions during reading. *Psychol Sci*. 25(6):1218–1226.
Schuler W, AbdelRahman S, Miller T, Schwartz L. 2010. Broad-coverage parsing using human-like memory constraints. *Comput Linguist*. 36(1):1–30.
Scott TL, Gallée J, Fedorenko E. 2017. A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cogn Neurosci*. 8(3):167–176.
Shain C. 2019. A large-scale study of the effects of word frequency and predictability in naturalistic reading. *NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.
Shain C, Blank IA, van Schijndel M, Schuler W, Fedorenko E. 2019. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *BioRxiv*. 717512.
Shain C, Blank IA, van Schijndel M, Schuler W, Fedorenko E. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*. 138:717512.
Shain C, Schuler W. 2019. Continuous-time Deconvolutional regression for psycholinguistic Modeling. *PsyArXiv*.
Shashidhara S, Spronkers FS, Erez Y. 2019. Individual-subject functional localization increases univariate activation but not multivariate pattern discriminability in the "multiple-demand" frontoparietal network. *J Cogn Neurosci*. 32(7):1348–1368.
Slevc LR, Rosenberg JC, Patel AD. 2009. Making psycholinguistics musical: self-paced reading time evidence for shared processing of linguistic and musical syntax. *Psychon Bull Rev*. 16(2):374–381.
Small SL, Nusbaum HC. 2004. On the neurobiological investigation of language understanding in context. *Brain Lang*. 89(2):300–311.
Smith NJ, Levy R. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*. 128(3):302–319.
Snijders TM, Vosse T, Kempen G, Van Berkum JJA, Petersson KM, Hagoort P. 2009. Retrieval and unification of syntactic structure in sentence comprehension: an fMRI study using word-category ambiguity. *Cereb Cortex*. 19(7): 1493–1503.
Speer NK, Zacks JM, Reynolds JR. 2007. Human brain activity time-locked to narrative even boundaries. *Psychol Sci*. 18(5):449–455.

Speer Nicole K, Reynolds JR, Swallow KM, Zacks JM. 2009. Reading stories activates neural representations of visual and motor experiences. *Psychol Sci*. 20(8):989–999.

Sreenivasan KK, Curtis CE, D'Esposito M. 2014. Revisiting the role of persistent neural activity during working memory. *Trends Cogn Sci*. 18(2):82–89.

Staub A. 2020. Do effects of visual contrast and font difficulty on readers' eye movements interact with effects of word frequency or predictability? *J Exp Psychol Hum Percept Perform*. 46(11):1235–1251.

Staub A, Goddard K. 2019. The role of preview validity in predictability and frequency effects on eye movements in reading. *J Exp Psychol Learn Mem Cogn*. 45(1):110–127.

Stowe LA, Broere CAJ, Paans AMJ, Wijers AA, Mulder G, Vaalburg W, Zwarts F. 1998. Localizing components of a complex task: sentence processing and working memory. *Neuroreport*. 9(13):2995–2999.

Sturt P. 2007. Semantic re-interpretation and garden path recovery. *Cognition*. 105(2):477–488.

Tahmasebi AM, Davis MH, Wild CJ, Rodd JM, Hakyemez H, Abolmaesumi P, Johnsrude IS. 2012. Is the link between anatomical structure and function equally strong at all cognitive levels of processing? *Cereb Cortex*. 22(7):1593–1603.

Taylor JSH, Rastle K, Davis MH. 2014. Interpreting response time effects in functional imaging studies. *Neuroimage*. 99:419–433.

Tettamanti M, Weniger D. 2006. Broca's area: a supramodal hierarchical processor? *Cortex*. 42(4):491–494.

Thesen S, Heid O, Mueller E, Schad LR. 2000. Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magn Reson Med*. 44(3): 457–465.

Thompson-Schill SL, Bedny M, Goldberg RF. 2005. The frontal lobes and the regulation of mental activity. *Curr Opin Neurobiol*. 15(2):219–224.

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

van Schijndel M, Exley A, Schuler W. 2013. A model of language processing as hierarchic sequential prediction. *Top Cognit Sci*. 5(3):522–540.

Vandenberghe R, Nobre AC, Price CJ. 2002. The response of left temporal cortex to sentences. *J Cogn Neurosci*. 14(4): 550–560.

Vasishth S, von der Malsburg T, Engelmann F. 2013. What eye movements can tell us about sentence comprehension. *Wiley Interdisc Rev*. 4(2):125–134.

Vázquez-Rodríguez B, Suárez LE, Markello RD, Shafiei G, Paquola C, Hagmann P, Van Den Heuvel MP, Bernhardt BC, Spreng RN, Misic B. 2019. Gradients of structure–function tethering across neocortex. *Proc Natl Acad Sci U S A*. 116(42):21219–21227.

Vergauwe E, Barrouillet P, Camos V. 2010. Do mental processes share a domain-general resource? *Psychol Sci*. 21(3):384–390.

Wehbe L, Murphy B, Talukdar P, Fyshe A, Ramdas A, Mitchell T. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS One*. 9(11):e112575.

Weichwald S, Meyer T, Özdenizci O, Schölkopf B, Ball T, Grosse-Wentrup M. 2015. Causal interpretation rules for encoding and decoding models in neuroimaging. *Neuroimage*. 110:48–59.

Wernicke C. 1874. Der aphasische Symptomencomplex. Eine psychologische Studie auf anatomischer basis. [the aphasia symptom complex. A psychological study on an anatomical basis]. *Wernicke's Work Aphasia*.

Whitfield-Gabrieli S, Nieto-Castanon A. 2012. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect*. 2(3):125–141.

Whitney C, Huber W, Klann J, Weis S, Krach S, Kircher T. 2009. Neural correlates of narrative shifts during auditory story comprehension. *Neuroimage*. 47(1):360–366.

Wild CJ, Yusuf A, Wilson DE, Peelle JE, Davis MH, Johnsrude IS. 2012. Effortful listening: the processing of degraded speech depends critically on attention. *J Neurosci*. 32(40):14010–14021.

Willems RM, Frank SL, Nijhof AD, Hagoort P, Van Den Bosch A. 2016. Prediction during natural language comprehension. *Cereb Cortex*. 26(6):2506–2516.

Wright P, Randall B, Marslen-Wilson WD, Tyler LK. 2011. Dissociating linguistic and task-related activity in the left inferior frontal gyrus. *J Cogn Neurosci*. 23(2):404–413.

Wright R, Ward L. 2008. Eye movements and attention shifts. *Orienting of attention*.

Yarkoni T, Barch DM, Gray JR, Conturo TE, Braver TS. 2009. BOLD correlates of trial-by-trial reaction time variability in gray and white matter: a multi-study fMRI analysis. *PLoS One*. 4(1).