fMRI reveals language-specific predictive coding during naturalistic sentence comprehension

Cory Shain\textsuperscript{a,1,*}, Idan Asher Blank\textsuperscript{b,c,1}, Marten van Schijndel\textsuperscript{d}, William Schuler\textsuperscript{a,e}, Evelina Fedorenko\textsuperscript{e}

\textsuperscript{a} The Ohio State University, 43210, USA
\textsuperscript{b} University of California Los Angeles, 90024, USA
\textsuperscript{c} Massachusetts Institute of Technology, 02139, USA
\textsuperscript{d} Cornell University, 14853, USA
\textsuperscript{e} Massachusetts General Hospital, Program in Speech and Hearing Bioscience and Technology, 02115, USA

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Language
Predictive coding
Multiple demand network
Naturalistic
fMRI
Sentence processing
Syntactic structure
Surprisal

\textbf{ABSTRACT}

Much research in cognitive neuroscience supports prediction as a canonical computation of cognition across domains. Is such predictive coding implemented by feedback from higher-order domain-general circuits, or is it locally implemented in domain-specific circuits? What information sources are used to generate these predictions? This study addresses these two questions in the context of language processing. We present fMRI evidence from a naturalistic comprehension paradigm (1) that predictive coding in the brain’s response to language is domain-specific, and (2) that these predictions are sensitive both to local word co-occurrence patterns and to hierarchical structure. Using a recently developed continuous-time deconvolutional regression technique that supports data-driven hemodynamic response function discovery from continuous BOLD signal fluctuations in response to naturalistic stimuli, we found effects of prediction measures in the language network but not in the domain-general multiple-demand network, which supports executive control processes and has been previously implicated in language comprehension. Moreover, within the language network, surface-level and structural prediction effects were separable. The predictability effects in the language network were substantial, with the model capturing over 37% of explainable variance on held-out data. These findings indicate that human sentence processing mechanisms generate predictions about upcoming words using cognitive processes that are sensitive to hierarchical structure and specialized for language processing, rather than via feedback from high-level executive control mechanisms.

\section{1. Introduction}

The human brain is an efficient prediction engine (\textit{James, 1890}). Facilitation in processing expected information, as well as processing costs of violated expectations, have been reported in many domains. In the domain of language comprehension, various results show that listeners and readers actively predict upcoming linguistic material (e.g., Kutas and Hillyard, 1984; MacDonald \textit{et al.,} 1994; Tanenhaus \textit{et al.,} 1995; Rayner \textit{et al.,} 2004; Frank and Bod, 2011; Smith & Levy, 2011, 2013; Gagnepain \textit{et al.,} 2012; Staub and Benatar, 2013; Frank \textit{et al.,} 2015; Kuperberg and Jaeger, 2016). However, the cognitive and neural mechanisms that support predictive language processing are not well understood. Under one widely held view, predictive language processing is implemented by domain-general executive (inhibitory control and working memory) resources. This perspective receives support from numerous studies showing that prediction effects during language comprehension are absent or less pronounced for populations with reduced executive resources, such as children, older individuals, and non-native speakers (e.g., Federmeier \textit{et al.,} 2002; Federmeier and Kutas, 2005; Dagerman \textit{et al.,} 2006; Federmeier \textit{et al.,} 2010; Mani and Huetting, 2012; Wlotko and Federmeier, 2012; Martin \textit{et al.,} 2013; Kaan, 2014; Mitsugi and MacWhinney, 2016; Gambi \textit{et al.,} 2018; Payne and

\textsuperscript{*} Corresponding author.
\textsuperscript{1} co-first author.

\textit{https://doi.org/10.1016/j.neuropsychologia.2019.107307}

Received 12 March 2019; Received in revised form 2 December 2019; Accepted 13 December 2019

Available online 24 December 2019

0028-3932/© 2019 Elsevier Ltd. All rights reserved.
Federmann, 2018; cf. Dave et al., 2018; Havron et al., 2019). Furthermore, several neuroimaging studies have reported sensitivity to linguistic manipulations in what appear to be cortical regions thought to support domain-general executive function (e.g., Kaan and Swaab, 2002; Kuperberg et al., 2003; Novick et al., 2005; Rodd et al., 2005; Novais-Santos et al., 2007; January et al., 2009; Peele et al., 2009; Rogalsky and Hickok, 2011; Nieuwland et al., 2012; Wild et al., 2012; McMillan et al., 2012, 2013), suggesting that such regions may also be implicated in language processing, including perhaps prediction. These results have led some to conclude that predictive coding for language is implemented by domain-general executive control resources (Linck et al., 2014; Huettig and Mani, 2016; Pickering and Gambi, 2018; Strijkers et al., 2019).

However, this interpretation is subject to several objections. First, most prior work on linguistic prediction has relied on behavioral and electrophysiological measures which are well suited for identifying global response patterns but cannot spatially localize the source of these effects in the brain to a certain functional region or network (e.g., Mather et al., 2013). Second, the (alleged) between-population differences in prediction noted above are consistent with accounts that do not directly invoke executive resources, including (1) possible qualitative differences between populations in the kind of information that is being predicted and the consequent need for population-specific norms to detect prediction effects, or (2) differences in how often predictions are correct, which may modulate the likelihood of engaging in predictive behavior (see Ryskin et al., this issue, for discussion). And third, past studies that did employ neuroimaging tools with high spatial resolution and consequently reported linguistic prediction responses—typically neural response increases for violations of linguistic structure—localized to executive control regions (e.g., Newman et al., 2001; Kuperberg et al., 2003; Nieuwland et al., 2012; Schuster et al., 2016) may have been influenced by task artifacts; indeed, some have argued that artificially constructed laboratory stimuli and tasks increase general cognitive load in comparison to naturalistic language comprehension (e.g., Blanco-Elorrieta and Pylkkänen, 2017; Blank and Fedorenko, 2017; Campbell and Tyler, 2018; Webbe et al., submitted; Diachek et al., 2019). To ensure that findings from the laboratory paradigms truly reflect the cognitive phenomenon of interest, it is important to validate them in more naturalistic experimental settings that better approximate the typical conditions of human sentence comprehension (Hasson and Honey, 2012; Hasson et al., 2018).

Despite the growing number of fMRI studies of naturalistic language comprehension (e.g., Speer et al., 2007; Yarkoni et al., 2008; Speer et al., 2009; Whitney et al., 2012; Webbe et al., 2014; Hale et al., 2015; Henderson et al., 2015, 2016; Huth et al., 2016; Sood and Sereno, 2016; Brennan, 2016; Desai et al., 2016; de Heer et al., 2017, Dehghani et al., 2017; Bhattasali et al., 2018), only a handful have directly investigated effects of word predictability (Willems et al., 2015; Brennan et al., 2016; Henderson et al., 2016; Lopopolo et al., 2017; see Table 1 for summary), a well-established predictor of behavioral measures in naturalistic language comprehension (Demberg and Keller, 2008; Frank and Bod, 2011; Smith and Levy, 2013; van Schijndel and Schuler, 2015). These previous naturalistic studies of linguistic prediction effects in the brain—using estimates of prediction effort such as surprisal (Hale, 2001; Levy, 2008), i.e., the negative log probability of a word given its context, or entropy (Hale, 2006), i.e., an information-theoretic measure of the degree of constraint placed by the context on upcoming words—have yielded mixed results on the existence, type, and functional location of such effects. For example, of the lexicalized and unlexicalized (part-of-speech) bigram and trigram models of word surprisal explored in Brennan et al. (2016), only part-of-speech bigrams positively modulated neural responses in most regions of the functionally localized language network. Lexicalized bi- and trigrams and part-of-speech trigrams yielded generally null or negative results (16 out of 18 comparisons). By contrast, Willems et al. (2015) found lexicalized trigram effects in regions typically associated with language processing (e.g., anterior and posterior temporal lobe). In addition, Willems et al. (2015) and Lopopolo et al. (2017) found prediction effects in regions that are unlikely to be specialized for language processing, including (aggregating across both studies) the brain stem, amygdala, putamen, and hippocampus, as well as in superior frontal areas more typically associated with domain-general executive functions like self-awareness and coordination of the sensory system (Goldberg et al., 2006). It is therefore not yet clear whether predictive coding for language relies on domain-general mechanisms in addition to, or instead of, language-specific ones, especially in naturalistic contexts.

In addition to questions about the functional localization of linguistic prediction, substantial prior work has also investigated the structure of the predictive model, seeking to shed light on the nature of linguistic representations in the mind. Specifically, if effects from theoretical constructs—such as hierarchical natural language syntax—can be detected in online processing measures (whether behavioral or neural), this would constitute evidence that such constructs are present in human mental representations and used to comprehend language. In the particular case of hierarchical representations, their psychological reality is widely supported by behavioral and electrophysiological experiments using constructed stimuli (see Lewis and Phillips, 2015 for review), by production experiments (Momma & Ferreira, 2019), and by some behavioral (Roark et al., 2009; Fossum and Levy, 2012; van Schijndel and Schuler, 2015; Shain et al., 2016), electrophysiological (Brennan and Hale, 2019) and neuroimaging (Brennan et al., 2016) experiments using naturalistic stimuli. However, other naturalistic studies reported null or negative syntactic effects (Frank and Bod, 2011; van Schijndel and Schuler, 2013; Shain and Schuler, 2018 contra Shain et al., 2016), or mixed syntactic results within the same set of experiments (Demberg and Keller, 2008; Henderson et al., 2016), leading some to argue that the representations used for language comprehension (in the absence of task artifacts from constructed stimuli) contain little hierarchical structure (Frank and Christiansen, 2018). Furthermore, the few naturalistic fMRI studies that have explored structural prediction effects have yielded inconsistent localization of these effects. For example, Brennan et al. (2016) found context-free grammar surprisal effects throughout the functional language network except in inferior frontal gyrus, whereas inferior frontal gyrus is the only region in which Henderson et al. (2016) found such effects.

The current study thus used fMRI to determine whether signatures of predictive coding during naturalistic language comprehension—increased response to less predictable words, i.e. surprisal effects (e.g., Smith and Levy, 2013), based on either linear word sequences or hierarchical structures—are primarily evident in (1) the domain-specific, fronto-temporal language (LANG) network (Fedorenko et al., 2011), and/or (2) the domain-general, fronto-parietal multiple

<table>
<thead>
<tr>
<th>Study</th>
<th># Participants</th>
<th>Stimulus length</th>
<th>HRF model</th>
<th>Functional localization</th>
<th>Out-of-sample evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Willems et al. (2015)</td>
<td>24</td>
<td>19 min</td>
<td>Canonical model</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Brennan et al. (2016)</td>
<td>26</td>
<td>12 min</td>
<td>Canonical model</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Henderson et al. (2016)</td>
<td>40</td>
<td>22 paragraphs</td>
<td>Canonical model</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Lopopolo et al. (2017)</td>
<td>22</td>
<td>19 min</td>
<td>Canonical model</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Current study</td>
<td>78</td>
<td>13.5 min (avg per participant)</td>
<td>Data-driven (CDR)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
The MD network supports executive functions (e.g., inhibitory control, attentional selection, conflict resolution, maintenance and manipulation of task sets) across both linguistic and non-linguistic tasks (e.g., Duncan and Owen, 2000; Fedorenko et al., 2013; Hugdahl et al., 2015; for discussion, see: Fedorenko, 2014) and has been shown to be sensitive to surprising events (Corbetta and Shulman, 2002).

On the one hand, given that the language network plausibly stores linguistic knowledge, including the statistics of language input, it might directly carry out predictive processing. Such a result would align with a growing body of cognitive neuroscience research supporting prediction as a “canonical computation” (Keller and Mrsc-Flogl, 2018) locally implemented in domain-specific circuits (Montague et al., 1996; Rao and Ballard, 1999; Alink et al., 2010; Bubic et al., 2010; Bastos et al., 2012; Wacongne et al., 2011, 2012; Singer et al., 2013). This hypothesis is also supported by prior findings of linguistic prediction effects in portions of the language network (Bonhage et al., 2015; Willems et al., 2015; Brennan et al., 2016; Henderson et al., 2016; Lopopolo et al., 2017; Matchin et al., 2018).

On the other hand, given that the MD network has been argued to encode predictive signals across domains and relay them as feedback to other regions (Strange et al., 2005; Cristescu et al., 2006; Egner et al., 2008; Wacongne et al., 2011; Chao et al., 2018), it might be recruited to predict upcoming words and structures in language. There is an extensive literature on neural signatures of prediction, such as activity associated with prediction errors, in brain regions that appear to belong to the MD network, including bilateral areas in the dorsolateral pre-frontal cortex, the inferior frontal gyrus, the anterior cingulate cortex, and the parietal lobe (for a review, see Dehaene et al., 2015; for a meta-analysis, see D’Astolfo and Rief, 2017). These areas are sensitive to rule violations in non-linguistic sequences, including hierarchically structured ones, in different sensory domains (e.g., auditory and visual; Bekinschtein et al., 2009; Ahlheim et al., 2014; Uhrig et al., 2014; Wang et al., 2015; Wang et al., 2017; Chao et al., 2018). In addition, they are recruited during learning of structured sequences in the motor domain (Bischoff-Grethe et al., 2004; Eickhoff et al., 2010). Beyond representing deterministic rules, such regions are also engaged in probabilistic predictions (Strange et al., 2005; Meyniel and Dehaene, 2017). Such predictions can be based on either inferring a generative model underlying the input sequence (Gläscher et al., 2010; Schaprio et al., 2013) or on reward contingencies (Koch et al., 2008; Zarr and Brown, 2016; Alexander and Brown, 2018; for a review, see: Rushworth and Behrens, 2008).

There are two main hypotheses in the contemporary literature that link predictive processing in the MD network with increased activity to more surprising words. First, the MD network might provide additional resources (“cognitive juice”) to various cognitive processes, including language. Under this scenario, MD regions might “come to the rescue” of the language network when processing demands are increased, which would be the case when surprisal is higher. Indeed, prior work suggests that the MD network could be recruited when language processing becomes effortful, e.g., under acoustic (Adank, 2012; Hervais-Adelman et al., 2012; Wild et al., 2013; Scott and McGettigan, 2013; Vaden et al., 2013) or syntactic (Kuperberg et al., 2003; Nieuwland et al., 2012) noise; in healthy aging (for reviews, see Wingfield and Grossman, 2006; Shafo and Tyler, 2014); during recovery from aphasia (Brownsett et al., 2014; Geranmayeh et al., 2014, 2016, 2017; Meier et al., 2016; Sims et al., 2016; Hartwigsen, 2018); and in L2 processing and multi-lingual contexts (e.g., Wexler and Beker, 2003; Rüschemeyer et al., 2005; Yokoyama et al., 2006; de Bruin et al., 2014; Grant et al., 2015; Kim et al., 2016; for reviews, see Perani and Abutalebi, 2005; Sakai, 2005; Abutalebi, 2008; Kotz, 2009; Hervais-Adelman et al., 2011; Pliatsikas and Luk, 2016). Second, the MD network, especially in the prefrontal cortex, may construct abstract representations of context, which serve as working memory for guiding behavior (Alexander and Brown, 2018).

The main goal of such representations is to minimize prediction errors in other brain regions, so these representations are communicated in a top-down manner to the language network or other domain-specific networks (e.g., sensory areas). Such high-level, abstract predictive signals are potentially useful because they could perhaps “explain away” some more local prediction errors computed in the language network (e.g., in a sentence like “the cat that the dog chased on the balcony escaped”, the verb “escaped” might be unexpected based on the local context of the previous few words, but its occurrence could be explained away by a more global and abstract representation that looks farther into the past and predicts a verb for “the cat” in the main clause). In essence, then, signals from the MD network could bias representations in the language network in favor of the features that are most relevant in a given context (for a similar reasoning for sensory cortices, see Miller and Cohen, 2001; Sreenivasan et al., 2014; D’Esposito and Posdle, 2015). However, these higher-level predictions are still sometimes incorrect, and when errors propagate back to the MD network, its regions would be triggered to adjust their predictive model in order to minimize future errors. This “model revision” process may register as increased neural processing (Chao et al., 2018).

Prior BMI studies using hand-constructed sentences to probe effects of linguistic expectation have not yielded a clear answer as to the mechanisms—language-specific vs. domain-general—that support linguistic prediction. Numerous such studies have observed responses in areas of the language network to manipulations of word predictability (Kuperberg et al., 2000; Baumgaertner et al., 2002; Kiell et al., 2002; Friederici et al., 2003; Gold et al., 2006; Obleser et al., 2007; Dien et al., 2008; Obleser and Kotz, 2009; Bonhage et al., 2015; Schuster et al., 2016, 2019; Hartwigsen et al., 2017; Matchin et al., 2018). However, as discussed above, many studies have also reported linguistic prediction effects in frontal, parietal, and cingulate cortical regions typically associated with the MD network (Kuperberg et al., 2000; Baumgaertner et al., 2002; Gold et al., 2006; Bonhage et al., 2015; Hartwigsen et al., 2017), as well as in other parts of the brain like the fusiform gyrus (Kuperberg et al., 2000; Gold et al., 2006) and the cerebellum (Lesage et al., 2017). Although it is certainly possible that predictive coding for language is carried out by both the LANG and the MD networks, with additional contributions from other brain areas, it is important to ensure that the foregoing results are not due to task artifacts induced by the use of artificially constructed stimuli (see Discussion), through validation of these findings in more naturalistic comprehension conditions (Hasson et al., 2018).

To distinguish between hypotheses above in a naturalistic comprehension paradigm, we searched for neural responses in LANG vs. MD regions to the contextual predictability of words as estimated by two model implementations of surprisal: a surface-level 5-gram model, and a hierarchical probabilistic context-free grammar (PCFG) model. N-gram surprisal estimates are sensitive to word co-occurrence patterns but are limited in their ability to model hierarchical natural language syntax, since they contain no explicit representation of grammatical categories or syntactic composition and have limited memory for preceding words in the sentence (in our case, up to four preceding words). PCFG surprisal estimates, by contrast, are based on structured syntactic representations of the unfolding sentence but do not directly encode surface-level word co-occurrence patterns. Correlations between each of these two measures and human neural responses would shed light on the relative importance assigned to these two information sources (word co-occurrences and syntactic structures) in computing predictions about upcoming words. Although surprisal is not the only extant measure of linguistic prediction (others include PCFG entropy, Roark et al., 2009; entropy reduction, Hale, 2006; and successor surprisal, Klügel et al., 2006), surprisal has received extensive consideration in the experimental literature (e.g., Demberg and Keller, 2008; Frank and Bod, 2011; Fossum and Levy, 2012; Frank et al., 2015; van Schijndel and Schuler, 2015; Brennan et al., 2016; Henderson et al., 2016; Brennan and Hale, 2019; Shain, 2019). We did not consider these related measures in order to avoid excessive statistical comparisons.

Note that by estimating prediction effects using surprisal, we are
implicitly assuming a notion of linguistic prediction as a distributed *pre*-activation process, following e.g. Kuperberg and Jaeger (2016), rather than as an all-or-nothing commitment to a specific upcoming word. Thus, we are investigating the degree to which the statistics of the local lexical (n-gram) and structural (PCFG) linguistic context modulate neural recruitment during the sentence processing, and which cognitive systems (language-specific vs. domain-general) are subject to such modulation. We leave aside questions about the underlying mechanisms by which these effects arise: e.g. the extent to which they are “active” or “passive”, or the extent to which integrative structure-building operations (e.g. composing words into syntactic constituents, constructing dependencies, etc.) underlie the observed facilitation effects (Altmann, 1998; Hale, 2014) (See Discussion for elaboration on this point.)

To identify language-specific and domain-general cortical regions without having to rely on the problematic practice of reverse inference from anatomy to function (Poldrack, 2006, 2011, Fig. 1), we functionally defined the LANG and MD networks in each individual participant using an independent “localizer” task (Saxe et al., 2006; Fedorenko et al., 2010), and then examined the response of those functional regions to each estimate of surprisal during naturalistic comprehension. To foreshadow our results, we find significant independent effects of 5-gram and PCFG surprisal in LANG, but no such effects in MD, as well as significant differences in surprisal effect sizes between the two networks. This finding supports the hypothesis that predictive coding for language is primarily carried out by language-specialized rather than domain-general cortical circuits and exploits both surface-level and structural cues.

2. Materials and methods

2.1. General approach

Several features set the current study apart from prior cognitive neuroscience investigations of linguistic prediction during comprehension.

First, we used naturalistic language stimuli rather than controlled stimuli constructed for a particular experimental goal. Naturalistic stimuli improve ecological validity compared to isolated constructed stimuli that may introduce task artifacts and not generalize to everyday cognition (Demberg and Keller, 2008; Hasson and Honey, 2012; Richlan et al., 2013; Schuster et al., 2016; Campbell and Tyler, 2018). Minimizing such artifacts is crucial in studies of the MD network, which is highly sensitive to task variables (Miller and Cohen, 2001; Sreenivasan et al., 2014; D’Esposito and Postle, 2015; Diachek et al., 2019). Moreover, prior work indicates that naturalistic stimuli yield more reliable fMRIBOLD signals than artificial tasks (Hasson et al., 2010).

Second, we used participant-specific functional localization to identify regions of interest constituting the LANG and MD networks (Fedorenko et al., 2010). This approach is crucial because many functional regions do not exhibit a consistent mapping onto macro-anatomical landmarks (e.g., Frost and Goebel, 2012; Tahmasebi et al., 2012; Vázquez-Rodríguez et al., 2019), and this variability is especially problematic when functionally distinct regions lie in close proximity to one another, as is the case in the frontal (Amunts et al., 1999; Tomaiuolo et al., 1999), temporal (Jones and Powell, 1970; Gloor, 1997; Wise et al., 2001) and parietal (Caspers et al., 2006, 2008; Schepfians et al., 2008) lobes, which house the LANG and MD networks. Due to this inconsistent functional-to-anatomical mapping, a given stereotactic coordinate might belong to the LANG network in some participants but to the MD network in others, as is indeed the case in our sample (Fig. 1) (see also Fedorenko et al., 2012a; Blank et al., 2017; Fedorenko & Blank, in press). Such inter-individual variability severely compromises the validity of both anatomical localization (Juch et al., 2005; Poldrack, 2006; Fischl et al., 2007; Frost and Goebel, 2012; Tahmasebi et al., 2012) and group-based functional localization (Saxe et al., 2006; Fedorenko and Kanwisher, 2009): these approaches risk

![Fig. 1. Inter-individual variability in the mapping of function onto anatomy. Each column demonstrates variability in a different coordinate in MNI space, specified at the top (in mm). For each coordinate, sagittal T1 slices from four participants are shown, with the coordinate circled on each slice (participants differ across columns). In each case, the top two participants show a Sentences > Nonwords effect in this coordinate (colored in red-yellow), whereas the bottom two participants show the opposite, Nonwords > Sentences effect in this same coordinate (colored in green-blue). In all cases, the effect size of the circled coordinate is strong enough to be included among the participant-specific fROIs. Other voxels exhibiting strong contrast effects in the localizer task (namely, among the top 10% of voxels across the neocortical gray matter) are superimposed onto the anatomical slices, in color. Colorbars show p-values associated with each of the two localizer contrasts. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)](image-url)
both decreased sensitivity (i.e. failing to identify a functional region due to insufficient spatial alignment across participants) and decreased functional resolution (i.e. mistaking two functionally distinct regions for a single region due to apparent spatial overlap across the sample). In contrast, participant-specific functional localization allows us to pool data from a given functional region across participants even in the absence of perfect anatomical alignment, and is therefore better suited for the kind of questions we study here (Nieto-Castañón and Fedorenko, 2012). Both networks we probe here have been extensively functionally characterized in prior work, so responses to linguistic surprisal therein can be taken to index the engagement of linguistic processing mechanisms vs. domain-general executive mechanisms (e.g., Mather et al., 2013).

Third, we analyzed the BOLD signal times-series using a recently developed statistical framework—continuous-time deconvolutional regression (CDR; Shain and Schuler, 2018, 2019)—that is designed to overcome problems in hemodynamic response modeling presented by naturalistic experiments. Namely, the variable spacing of words in naturalistic language prevents direct application of discrete-time, data-driven techniques for hemodynamic response function (HRF) discovery, such as finite impulse response modeling (FIR) or vector autoregression. Because CDR is a parametric continuous-time deconvolutional method, it can infer the hemodynamic response directly from naturalistic time series, without distortionary preprocessing steps such as predictor interpolation (cf. Huth et al., 2016). Thus, unlike prior naturalistic fMRI studies of prediction effects in language processing (Table 1), we do not assume the shape of the HRF.

Fourth, unlike studies in Table 1, we evaluated hypotheses using non-parametric statistical tests of model fit to held-out (out-of-sample) data, an approach which builds external validity directly into the statistical test and should thereby improve replicability (e.g., Demšar, 2006).

Finally, to our knowledge, this is the largest fMRI investigation to date (78 participants) of prediction effects in naturalistic language comprehension.

3. Experimental design

3.1. Participants

Seventy-eight native English speakers (30 males), aged 18–60 (M±SD = 25.8 ± 9, Med ± IQR = 23 ± 3), from MIT and the surrounding Boston community participated for payment. Each participant completed a passive story comprehension task (the critical experiment) and a functional localizer task designed to identify the LANG and MD networks.

Sixty-nine participants (88%) were right-handed, as determined by either the Edinburgh handedness inventory (n = 66) (Oldfield, 1971) or self-report (n = 11) (handedness data were not collected for one participant). Eight participants were left-handed, but seven of these showed typical left-lateralized language activations, as determined by examining their activation patterns for the language localizer task (see below); the remaining participant had a right-lateralized language network. We chose to include the latter participant’s data in the analyses, to err on the conservative side, and to be able to generalize the results to the population at large (see Willems et al., 2015; for discussion).

All participants gave informed consent in accordance with the requirements of MIT’s Committee on the Use of Humans as Experimental Subjects (COUHES).

3.2. Stimuli and procedure

The localizer task and critical (story comprehension) experiment were run either in the same scanning session (67 participants) or in two separate sessions (11 participants, who have performed the localizer task while participating in other studies; see Mahowald and Fedorenko, 2016, for evidence of high stability of language localizer activations across sessions). For the critical experiment, each participant listened to one or more stories (one story: n = 34; two stories: n = 14; three stories: n = 13; four stories: n = 2; five stories: n = 4; six stories: n = 5; seven stories: n = 1; or eight stories: n = 5). In each session, participants performed a few other, unrelated tasks, with scanning sessions lasting approximately 2 h.

Localizer task. We used a single localizer task to identify functional regions of interest in both the LANG and MD networks, using opposite task contrasts across these networks as described below. This task, which has been described in more detail elsewhere (Fedorenko et al., 2010), consisted of reading sentences and lists of unconnected, pronounceable nonwords in a standard two-condition blocked design with a counter-balanced order across runs. Stimuli were presented one word/nonword at a time. The majority of participants (n = 60) read these materials passively (and pressed a button at the end of each trial, to sustain alertness); for some participants (n = 18), every trial ended with a memory probe item, and they had to indicate via a button press whether or not this probe had appeared in the preceding sentence/nonwords sequence. In addition, different participants performed versions of the task differing slightly in stimulus timing, number of blocks, etc., i.e. features that do not affect the robustness of the contrast (e.g., Fedorenko et al., 2010; Mahowald and Fedorenko, 2016) (for experimental parameters, see Table 2). A version of this localizer is available at https://evlab.mit.edu/funcloc/download-paradigms.

To identify LANG regions, we used the contrast sentences > nonwords. This contrast targets higher-level aspects of language, to the exclusion of perceptual (speech/reading) and motor-articulatory processes (for discussion, see Fedorenko and Thompson-Schill, 2014; or Fedorenko, in press). Critically, this localizer has been extensively validated over the past decade across diverse parameters: it generalizes across tasks (passive reading vs. memory probe), presentation modalities (visual vs. auditory), and task contrasts across these networks as described below. This task, which has been described in more detail elsewhere (Fedorenko et al., 2010), consisted of reading sentences and lists of unconnected, pronounceable nonwords in a standard two-condition blocked design with a counter-balanced order across runs. Stimuli were presented one word/nonword at a time. The majority of participants (n = 60) read these materials passively (and pressed a button at the end of each trial, to sustain alertness); for some participants (n = 18), every trial ended with a memory probe item, and they had to indicate via a button press whether or not this probe had appeared in the preceding sentence/nonwords sequence. In addition, different participants performed versions of the task differing slightly in stimulus timing, number of blocks, etc., i.e. features that do not affect the robustness of the contrast (e.g., Fedorenko et al., 2010; Mahowald and Fedorenko, 2016) (for experimental parameters, see Table 2). A version of this localizer is available at https://e

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Experimental parameters for the different versions of the localizer task.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version</td>
<td>A</td>
</tr>
<tr>
<td>Number of participants</td>
<td>60</td>
</tr>
<tr>
<td>Task (Passive Reading/Memory)</td>
<td>PR</td>
</tr>
<tr>
<td>Words/nonwords per trial</td>
<td>12</td>
</tr>
<tr>
<td>Trial duration (ms)</td>
<td>6000</td>
</tr>
<tr>
<td>Fixation</td>
<td>100</td>
</tr>
<tr>
<td>Presentation of each word/nonword</td>
<td>450</td>
</tr>
<tr>
<td>Probe (M) + button press (M/PR)</td>
<td>400</td>
</tr>
<tr>
<td>Trials per block</td>
<td>3</td>
</tr>
<tr>
<td>Blocks per condition (per run)</td>
<td>8</td>
</tr>
</tbody>
</table>

Notes: Conditions: Sentences Nonwords Sentences Nonwords Word-lists Sentences Nonwords Word-lists

Fixation block duration (s) 14 18 16 18
Number of fixation blocks 5 5 3 4
Total run time (s) 358 378 336 396
Number of runs 2 2 3 4

* Used for the purposes of another experiment; see (Fedorenko et al., 2010).
auditory), and materials (e.g., Fedorenko et al., 2010; Brazel et al., 2011; Vaghaarshakian et al., 2012), including both coarser contrasts (e.g., between natural speech and an acoustically degraded control: Scott et al., 2017) and narrower contrasts (e.g., between lists of unconnected, real words and nonword lists, or between sentences and lists of real words: Fedorenko et al., 2010; Blank et al., 2016). Whereas there are many potential differences (linguistic and otherwise) between the processing of sentences vs. nonwords, all regions localized with the sentences > nonwords contrast show a similar response profile: on the one hand, they exhibit sensitivity to various aspects of linguistic processing, including (but not limited to) lexical, phrasal, and sentence-level semantic and syntactic processing (e.g., Fedorenko et al., 2012b, 2018; Blank et al., 2016; Mollica et al., 2018; Blank & Fedorenko, in revision; similar patterns obtain in electrocorticographic data with high temporal resolution: Fedorenko et al., 2016). On the other hand, they show robust language-selectivity in their responses, with little or no response to non-linguistic tasks, including domain-general contrasts targeting, e.g., working memory or inhibitory control (Fedorenko et al., 2011, 2012a).

In other words, the localizer shows both convergent construct validity and a broad sensitivity to various aspects of linguistic processing, including (but not limited to) lexical, phrasal, and sentence-level semantic and syntactic processing (e.g., Fedorenko et al., 2012b, 2018; Blank et al., 2016; Mollica et al., 2018; Blank & Fedorenko, in revision; similar patterns obtain in electrocorticographic data with high temporal resolution: Fedorenko et al., 2016). On the other hand, they show robust language-selectivity in their responses, with little or no response to non-linguistic tasks, including domain-general contrasts targeting, e.g., working memory or inhibitory control (Fedorenko et al., 2011, 2012a).

To identify MD regions, we used the sentences > nonwords contrast, targeting regions that increase their response with the more effortful reading of nonwords compared to that of sentences. This “cognitive effort” contrast robustly engages the MD network and can reliably localize it. Moreover, it generalizes across a wide array of stimuli and tasks, both linguistic and non-linguistic including, critically, contrasts targeting executive functions such as working-memory and inhibitory control (Fedorenko et al., 2013; Mineroff et al., 2018). Supplementary Figs. 1 and 2 demonstrate that the MD regions thus localized robustly respond to a difficulty (i.e. memory load) manipulation in a non-linguistic, spatial working-memory task (administered to a subset of participants in the current dataset).

Main (story comprehension) task. Participants listened to stories from the publicly available Natural Stories Corpus (Futrell et al., 2018). These stories were adapted from existing texts (fairey tales and short stories) to be “deceptively naturalistic”: they contained an over-representation of rare words and syntactic constructions embedded in otherwise natural linguistic context. Behavioral data indicate that these stories effectively manipulate predictive processing, as self-paced reading times from an independent sample show robust effects of surprise (Futrell et al., 2018). Stories were recorded by two native English speakers (one male, one female) at a 44.1 kHz sampling rate, ranged in length from 4m46s to 6m29s (983–1099 words), and were played over scanner-safe headphones (Sensometrics, Malden, MA).

Following each story, some participants answered six (n = 29) or twelve (n = 12) comprehension questions, presented in a two-alternative forced-choice format. For all but 4 of these participants, accuracy was significantly above chance (nominal test for each participant: all ps < 0.046, uncorrected). For the remaining participants, comprehension questions were not part of the experimental design (n = 30), were not collected due to equipment malfunction (n = 4), or were lost (n = 3). We note that BOLD signal time-series show indistinguishable levels of stimulus-locked activity regardless of whether comprehension questions are administered or not, at least in the networks studied here (Blank and Fedorenko, 2017).

3.3. Data acquisition and preprocessing

Data acquisition. Structural and functional data were collected on a whole-body 3 Tesla Siemens Trio scanner with a 32-channel head coil at the Athinoula A. Martins Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted structural images were collected in 176 axial slices with 1 mm isotropic voxels (repetition time (TR) = 2,530 ms; echo time (TE) = 3.48 ms). Functional, blood oxygenation level-dependent (BOLD) data were acquired using an EPI (echo-planar imaging) sequence with a 90° flip angle and using GRAPPA (Generalized Autocalibrating Partial Parallel Acquisition) with an acceleration factor of 2; the following parameters were used: thirty-one 4.4 mm thick near-axial slices acquired in an interleaved order (with 10% distance factor), with an in-plane resolution of 2.1 mm × 2.1 mm, FoV (field of view) in the phase encoding (Anterior – Posterior) direction 200 mm and matrix size 96 mm × 96 mm, TR = 2000 ms and TE = 30 ms. The first 10 s of each run were excluded to allow for steady state magnetization.

Spatial preprocessing. Data preprocessing was carried out with SPM5 and custom MATLAB scripts. 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 200s. Note that SPM was only used for preprocessing and basic first-level modeling, aspects that have not changed much in later versions; 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.

Temporal preprocessing. Data from the story comprehension runs were additionally preprocessed 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-series from the WM were regressed out of each voxel’s time-course; signal originating in the CSF was similarly regressed out. Six principal components of the six motion parameters estimated during offline motion correction were also regressed out, as well as their first time derivative.

3.4. Participant-specific functional localization of the LANG and MD networks

Modeling localizer data. A general linear model estimated the voxel-wise effect size of each condition in each experimental run of the localizer task. These effects were each modeled with a boxcar function (representing entire blocks/events) convolved with the canonical Hemodynamic Response Function (HRF). The model also included first-order temporal derivatives of these effects, as well as nuisance regressors representing entire experimental runs and offline-estimated motion parameters. The obtained beta weights were then used to compute the two functional contrasts of interest: sentences > nonwords for identifying LANG regions, and nonwords > sentences for identifying MD regions. These contrasts were computed only for voxels whose probability of belonging to the gray matter was greater than 1/3, based on the segmentation of the participant’s anatomical data. All other voxels were not considered further.

Defining functional regions of interest (fROIs). For each participant, functional ROLs were defined by combining two sources of information (Fedorenko et al., 2010; Julian et al., 2012): (i) the participant’s activation map for the relevant localizer contrast (converted from beta weights to t-scores), and (ii) group-level constraints (”masks” available for download from https://evlab.mit.edu/functoc/download-parcels). The latter demarcated brain areas within which most or all individuals in prior studies showed activity for the localizer contrasts (Fig. 2).
For the LANG fROIs, we used masks derived from a group-level probabilistic representation of the sentences > nonwords contrast in a set of 220 participants. These masks 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, temporal, and parietal cortices. In particular, six masks were used: in the inferior frontal gyrus (IFG) and its orbital part (IFGorb), middle frontal gyrus (MFG), anterior temporal cortex (AntTemp), posterior temporal cortex (PostTemp), and angular gyrus (AngG).

For the MD fROIs, we used masks derived from a group-level probabilistic representation of data from a previously validated MD-localizer task in a set of 197 participants. The task, described in detail in Fedorenko et al. (2011), contrasted hard and easy versions of a visuo-spatial working memory task (we did not use masks based on the nonwords > sentences contrast in order to maintain consistency with other current projects in our lab, and because prior work has established the similarity of the activation landscapes for these two contrasts; Fedorenko et al., 2013). These masks were constrained to be bilaterally symmetric by averaging individual hard > easy contrast maps across the two hemispheres prior to generating the group-level representation (only the group-based masks, covering large swaths of cortex, were constrained in this way; fROIs in the current study were free to vary in their location across hemispheres, within the borders of these masks). The topography of these masks largely overlapped with anatomically based masks that we had used in previous work (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 (PpreG), superior frontal gyrus (SFG), middle frontal gyrus (MFG) and its orbital part (MF Görb), opercular part of the inferior frontal gyrus (IFGop), the anterior cingulate cortex and pre-supplementary motor cortex (ACC/pSMA), and the insula (Insula).

These group-level masks, in the form of binary maps, were used to constrain the selection of participant-specific fROIs. In particular, for each participant, 6 LANG fROIs were created by (i) intersecting each LANG mask with each individual participant’s unthresholded t-map for the sentences > nonwords contrast; and then (ii) choosing the 10% of voxels with highest t-scores in the intersection. Similarly, 20 MD fROIs were created by intersecting each MD mask with each participant’s unthresholded t-map for the nonwords > sentences contrast and selecting the 10% of voxels with the highest t-scores within each intersection. This top-10% criterion balances the trade-off between choosing only voxels that respond robustly to the relevant contrast and having a sufficient number of voxels in each fROI of each participant. Moreover, this criterion guarantees fROIs of identical size across participants (occupying 10% of each mask). Few exceptions to this criterion were made for those cases where less than 10% of the voxels in a mask showed a t-score greater than 0; here, we only included the subset of voxels with positive t-scores in the fROI, and excluded those voxels showing effects in the opposite direction.

Prior to the critical statistical analyses, we ensured that all fROIs showed the expected functional signatures, i.e. a sentences > nonwords effect for the LANG fROIs, and a nonwords > sentences effect for the MD fROIs. To this end, the reliability of each contrast effect (i.e. the difference between the beta estimates of the two localizer conditions) was tested using a 2-fold across-run cross-validation: for each participant, fROIs were defined based on odd (even) run(s) and, subsequently, independent estimates of the relevant contrast effect were obtained from the left-out even (odd) run(s). These contrast effects were averaged across the two partitions (odd/even) and tested for significance across participants, via a dependent samples t-test (FDR-corrected for the number of fROIs within each network). The sentence > nonwords effect was highly reliable throughout the language network (for all six fROIs: \( t_{177} > 9.5, p < 10^{-12} \) corrected; conservative effect size based on an independent samples test: Cohen’s \( d > 0.82 \), and the nonwords > sentences effect was highly reliable throughout the MD network (for all 20 fROIs: \( t_{177} > 2.25, p < 0.05 \); conservative effect size based on an independent samples test: Cohen’s \( d > 0.16 \) (see also Supplementary Figs. 1 and 2 for evidence of overlap with a spatial working memory contrast, as in Fedorenko et al., 2013).
4. Statistical analysis

4.1. Predictor definitions

To estimate word predictability in naturalistic data, we used an information-theoretic measure known as surprisal (Shannon, 1948; Hale, 2001): the negative log probability of a word given its context. Surprisal can be computed in many ways, depending on the choice of probability model. Three previous naturalistic fMRI studies (Willems et al., 2015; Brennan et al., 2016; Lopopolo et al., 2017) searched for surface-level n-gram surprisal effects, using words and/or parts of speech as the token-level representation. In addition, two previous naturalistic fMRI studies (Brennan et al., 2016; Henderson et al., 2016) probed structure-sensitive PCFG surprisal measures (Hale, 2001; Roark et al., 2009). As discussed in the Introduction, results from these studies failed to converge on a clear answer as to the nature and functional location of surprisal effects. In this study, we therefore used the following surprisal estimates:

- **5-gramSurprisal**: 5-gram surprisal for each word in the stimulus set from a KenLM (Heafield et al., 2013) language model with default smoothing parameters trained on the Gigaword 3 corpus (Graff et al., 2007). 5-gram surprisal quantifies the predictability of words as the negative log probability of a word given the four words preceding it in context.
- **PCFG Surprisal**: Lexicalized probabilistic context-free grammar surprisal computed using the incremental left-corner parser of van Schijndel et al. (2013) trained on a generalized categorial grammar (Nguyen et al., 2012) reannotation of Wall Street Journal sections 2 through 21 of the Penn Treebank (Marcus et al., 1993).

Models also included the control variables **Sound Power, Repetition Time (TR) Number, Rate, Frequency, and Network**, which were operationalized as follows:

- **Sound Power**: Frame-by-frame root mean squared energy (RMSE) of the audio stimuli computed using the Librosa software library (McFee et al., 2015).
- **TR Number**: Integer index of the current fMRI sample within the current scan.
- **Rate**: Deconvolutional intercept. A vector of ones time-aligned with the word onsets of the audio stimuli. Rate captures influences of stimulus timing independently of stimulus properties (see e.g., Brennan et al., 2016; Shain and Schuler, 2018).
- **Frequency**: Corpus frequency computed using a KenLM unigram model trained on Gigaword 3. For ease of comparison to surprisal, frequency is represented here on a surprisal scale (negative log probability), such that larger values index less frequent words (and thus greater expected processing cost).
- **Network**: Numeric predictor for network ID, 0 for MD and 1 for LANG.

Models additionally included the mixed-effects random grouping factors **Participant** and **JROI**. Prior to regression, all predictors were rescaled by their standard deviations in the training set except Rate (which has no variance) and Network (which is an indicator variable). Reported effect sizes are therefore in standard units.

4.2. Continuous-time deconvolutional regression

Naturalistic language stimuli pose a challenge for established statistical methods in fMRI because the stimuli (words) (1) are variably spaced in time and (2) do not temporally align with response samples recorded by the scanner. Previous approaches to address this issue have various drawbacks. Some fMRI studies of naturalistic language processing have assumed a canonical shape for the hemodynamic response function (Boynton et al., 1996) and used it to convolve stimulus properties into response-aligned measures (Willems et al., 2015; Brennan et al., 2016; Lopopolo et al., 2017). This approach is unable to account for regional variation in the shape of the hemodynamic response, even though the canonical HRF is known to be a poor fit to some brain regions (Handwerker et al., 2004). Discrete-time methods for data-driven HRF identification such as finite impulse response modeling (FIR; Duyal and MacGregor, 1996) and vector autoregression (VAR; Sims, 1980) are widely used to overcome the limitations of the canonical HRF for fMRI research (e.g., Friston et al., 1994; Harrison et al., 2003) but are of limited use in the naturalistic setting because they assume (multiples of) a fixed time interval between stimuli that does not apply to words in naturally-occurring speech. Some studies (e.g., Huth et al., 2016) address this problem by continuously interpolating word properties, resampling the interpolated signal so that it temporally aligns with the fMRI record, and fitting FIR models using the resampled design matrix. However, this approach can be distortionary in that word properties (e.g., surprisal) are not temporally continuous.

Our study employed a recently developed continuous-time deconvolutional regression (CDR) technique that accurately infers parametric continuous-time impulse response functions—such as the HRF—from arbitrary time series (Shain and Schuler, 2018, 2019). Because CDR is data-driven, it can address the potential impact of poor fit in the canonical HRF, and because it is defined in continuous time, it eliminates the need for distortionary preprocessing steps like continuous interpolation. CDR models in this study used the following two-parameter HRF kernel based on the widely-used double-gamma canonical HRF (Lindquist et al., 2009):

\[
h(x; \alpha, \beta) = \frac{\beta^6 e^{-x + \beta \alpha}}{T(\alpha)} - \frac{1}{6} \frac{\beta^6 e^{-10x - \beta \alpha + 9\alpha}}{T(\alpha + 10)}
\]

where \(\alpha\) and \(\beta\) are initialized to the SPM defaults of 6 and 1, respectively. More complex kernels (e.g., that fit the amplitude of the second term, rather than fixing it at 1/6) were avoided because of their potential to overfit.

The parametric continuous-time nature of CDR is similar to that of models used, for example, by Kruggel and von Cramon (1999), Kruggel et al. (2000), Miezin et al. (2000), Lindquist and Wager (2007), and Lindquist et al. (2009) for nonlinear estimation of gamma-shaped HRFs. The main advantages of CDR over these approaches are that (1) it exploits the Tensorflow (Abadi et al., 2015) and Edward (Tran et al., 2016) libraries for optimizing large-scale variational Bayesian computation graphs using state of the art estimation techniques from deep learning—this study used the Adam optimizer with Nesterov momentum (Kingma and Ba, 2014; Nesterov, 1983; Dozat, 2016); (2) it supports mixed effects modeling of effect coefficients and HRF parameters; and (3) it supports parameter tying, constraining the solution space by ensuring that all predictors share a common HRF shape in a given region (with potentially differing amplitudes). Predictors in these models were given their own coefficients (which rescale \(h\) above), but the parameters \(\alpha\) and \(\beta\) of \(h\) were tied across predictors, modeling the assumption of a fixed-shape blood oxygenation response to neural activity in a given cortical region.

The CDR models applied in this study assumed improper uniform priors over all parameters in the variational posterior and were optimized using a learning rate of 0.001 and stochastic minibatches of size 1024. Following standard practice from linear mixed-effects regression (Bates et al., 2015), random effects were L2-regularized toward zero at a rate of 1.0. Convergence was declared when the loss was uncorrelated with training time by t-test at the 0.5 level for at least 250 of the past 500 training epochs. For computational efficiency, predictor histories were truncated at 256 timesteps (words), which yields a maximum temporal coverage in our data of 48.34s (substantially longer than the effective influence of the canonical HRF). Prediction from the network used an exponential moving average of parameter iterates (Polyak and Juditsky,
with a decay rate of 0.999, and models were evaluated using maximum a posteriori estimates obtained by setting all parameters in the variational posterior to their means. This approach is valid because all parameters are independent Gaussian in the CDR variational posterior (Shain and Schuler, 2018).

4.3. Model specification

The following CDR model specification was fitted to responses from each of the LANG and MD fROIs, where italics indicate predictors convolved using the fitted HRF and bold indicates predictors that were ablated for hypothesis tests:

$\text{BOLD} \sim \text{TRNumber} + \text{soundPower} + \text{Rate} + \text{Frequency} + 5\text{-gram} + \text{PCFG} + (\text{TRNumber} + \text{soundPower} + \text{Rate} + \text{Frequency} + 5\text{-gram} + \text{PCFG} | \text{fROI}) + (1 \mid \text{Participant})$.

The random effect by fROI indicates that the model included zero-centered by-fROI random variation in response amplitude and HRF parameters for each functional region of interest. As shown, the model also included a random intercept by participant (the data do not appear to support richer by-participant random effects, e.g. including random slopes and HRF shapes, since such models explained no held-out variance in early analyses, indicating overfitting). The above model can test whether the surprisal variables help predict neural activation in a given cortical region. However, it cannot be used to compare the magnitudes of response to surprisal across networks (Nieuwenhuis et al., 2011).

Therefore, we directly tested for a difference in influence by fitting the combined responses from both LANG and MD using the following model specification with the indicator variable Network:

$\text{BOLD} \sim \text{TRNumber} + \text{soundPower} + \text{Rate} + \text{Frequency} + 5\text{-gram} + \text{PCFG} + \text{Network} + \text{TRNumber:Network} + \text{soundPower:Network} + \text{Rate:Network} + \text{Frequency:Network} + 5\text{-gram:Network} + \text{PCFG:Network} + (1 \mid \text{fROI}) + (1 \mid \text{Participant})$.

The random effects by fROI were simplified in comparison to that of the single-network models because the Network variable exactly partitions the fROIs. Thus ablated models can fully capture network differences as long as they have by-fROI random effects for surprisal. Indeed, initial tests showed virtually no difference in held-out likelihood between full and ablated combined models when those models included full by-fROI random effects despite large-magnitude estimates for the interactions with Network in the full model. Furthermore, the fitted parameters suggested that the by-fROI term was being appropriated in ablated models to capture between-network differences. In the full model, the 5-gram Surprisal estimates for 50% of LANG fROIs and 45% of MD fROIs were positive, while in the model with 5-gram:Network ablated, 100% of LANG fROIs and only 20% of MD fROIs were positive, indicating that differences in response to 5-gram Surprisal had been ‘pushed’ into the by-fROI random term. For this reason, we used simpler models for the combined test, despite their insensitivity to by-fROI variation in HRF shape or response amplitude.

In interactions between Network and convolved predictors, the interaction was computed following convolution but prior to rescaling with that predictor’s coefficient. Thus, the interaction term represents the offset in the estimated coefficient from the MD network to the LANG network, as is the case for binary interaction terms in linear regression models.

Finally, exact deconvolution from continuous predictors like Sound Power is not possible, since such predictors do not have an analytical form that can be integrated. Instead, we sampled sound power at fixed intervals (100 ms), in which case the event-based CDR procedure reduces to a Riemann sum approximation of the continuous convolution integral. Note that the word-aligned predictors (e.g. 5-gram Surprisal) therefore have different timestamps than Sound Power, and as a result the history window spans different regions of time (up to 128 words into the past for the word-aligned predictors and up to 100 ms × 128 = 12.8s of previous Sound Power samples).

4.4. Ablative statistical testing

In order to avoid confounds from (1) collinearity in the predictors and/or (2) overfitting to the training data, we followed a standard testing protocol from machine learning of evaluating differences in prediction performance on out-of-sample data using ablative non-parametric paired permutation tests for significance (Demsar, 2006). This approach can be used to assess the presence of an effect by comparing the prediction performance of a model that contains the effect against that of an ablated model that does not contain it. Specifically, given two pre-trained nested models, we computed the out-of-sample by-item likelihoods from each model over the evaluation set and constructed an empirical p value for the likelihood difference test statistic by randomly swapping by-item likelihoods n times (where n = 10,000) and computing the proportion of obtained likelihood differences whose magnitude exceeded that observed between the two models. To ensure a single degree of freedom for each comparison, only fixed effects were ablated, with all random effects retained in all models.

The data partition was created by cycling TR numbers e into different bins of the partition with a different phase for each participant $\pi$:

$\text{partition}(e; \pi) = \frac{e + \mu}{30} \mod 2$.

assigning output 0 to the training set and 1 to the evaluation set. Since TR duration is 2s, this procedure splits the BOLD times series into 60 s chunks, alternating assignment of chunks into training and evaluation sets with a different phase for each participant. Partitioning in this way allowed us to (1) obtain a model of each participant, (2) cover the entire time series, and (3) sub-sample different parts of the time series for each participant during training, while at the same time suppressing correlation between the training and evaluation responses by using a relatively long period of alternation (30 TRs or 60s).

4.5. Accessibility

Access instructions for software and supplementary data needed to replicate these experiments (e.g. librosa, PyMVPA, CDR, KenLM, Giga-word 3, etc.) are given in the publications cited above. Post-processed fMRI timeseries are publicly available at the following URL: http://ps://ofsl.io/eyrp8q/. These experiments were not pre-registered.

5. Results

The CDR-estimated mean double-gamma hemodynamic response functions (HRFs) for the LANG and MD networks are given in Fig. 3, the estimated HRFs by fROI in LANG regions are shown in Fig. 4, surprisal estimates and percent variance explained by fROI are given in Tables 3 and 4, and population-level effect estimates (i.e. areas under the estimated HRFs) are reported in Table 5. MD estimates by region are plotted in Supplementary Figs. 3 and 4; they are of little relevance because they do not generalize (Tables 4 and 6). As shown, HRF shapes resemble but deviate slightly from the canonical HRF (Boynton et al., 1996) to varying degrees in each region, highlighting both consistency with HRF estimates established by prior research as well as the potential of CDR to discover subtle differences in HRF shape between cortical regions (Handwerker et al., 2004) in naturalistic data. The models find positive effects of similar strength for both 5-gram Surprisal and PCFG Surprisal in LANG, and smaller effects of surprisal (even negative in the case of 5-gram Surprisal) in MD.

At the level of individual regions, the models explained held-out variance in all but one of the LANG fROIs (the exception was the AngG fROI). In contrast, the models explained no held-out variance in any but one MD fROI (the left MFGorb fROI). We leave these two exceptions to future research, but overall, the results demonstrate that...
surprisal effects are generally present throughout the LANG network and generally absent throughout the MD network. The differences between the individual-network models are largely replicated in the Combined model (Table 5), where main effects represent the estimated mean response in MD while interactions with Network represent the estimated difference in mean response between LANG and MD. As shown, Combined model estimates of both 5-gram:Network and PCFG:Network are positive and large-magnitude, indicating that the model estimates these variables to yield greater increases in neural activity in LANG over MD.

Table 3
LANG surprisal estimates by fROI. Estimates given are the area under the fitted HRF. Models explain held-out variance in all regions but AngG.

<table>
<thead>
<tr>
<th>ROI</th>
<th>Hemisphere</th>
<th>5-gram estimate</th>
<th>PCFG estimate</th>
<th>% Held-Out Variance Explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>AngG</td>
<td>L</td>
<td>0.030</td>
<td>0.156</td>
<td>0.0%</td>
</tr>
<tr>
<td>AntTemp</td>
<td>L</td>
<td>0.215</td>
<td>0.017</td>
<td>5.1%</td>
</tr>
<tr>
<td>IFG</td>
<td>L</td>
<td>0.287</td>
<td>0.309</td>
<td>2.2%</td>
</tr>
<tr>
<td>IFGorb</td>
<td>L</td>
<td>0.010</td>
<td>0.318</td>
<td>1.3%</td>
</tr>
<tr>
<td>MFG</td>
<td>L</td>
<td>0.382</td>
<td>0.346</td>
<td>2.3%</td>
</tr>
<tr>
<td>PostTemp</td>
<td>L</td>
<td>0.242</td>
<td>0.258</td>
<td>6.1%</td>
</tr>
</tbody>
</table>

Fig. 3. Estimated overall double-gamma hemodynamic response functions (HRFs) by network. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Fig. 4. Estimated language-network HRFs by fROI.

Table 6 reports model percent variance explained compared to a theoretical ceiling computed by regressing individual participants’
amplitude is not straightforward, and systematic investigation of re-
substantial amount of the available variance in LANG. MD models
mates suggests potentially interesting functional differences within the
future work. However, a qualitative examination of the by-region esti-
gions/subnetworks within each broader functional network is left to
pairwise statistical testing of between-region differences in effect
explain no variance on the evaluation set, suggesting that the MD model
pendently of inter-participant variation. As shown, models explain a

Table 4
MD surprisal estimates by fROI. Estimates given are the area under the fitted
HRF. Models explain no held-out variance in any region except left MFGorb.

<table>
<thead>
<tr>
<th>fROI</th>
<th>Hemisphere</th>
<th>5-gram Surprisal estimate</th>
<th>PCFG Surprisal estimate</th>
<th>% Held-Out Variance Explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>AntPar</td>
<td>L</td>
<td>0.102</td>
<td>−0.523</td>
<td>0.0%</td>
</tr>
<tr>
<td>IFGpop</td>
<td>L</td>
<td>0.009</td>
<td>0.141</td>
<td>0.0%</td>
</tr>
<tr>
<td>Insula</td>
<td>L</td>
<td>−0.200</td>
<td>0.284</td>
<td>0.0%</td>
</tr>
<tr>
<td>MFG</td>
<td>L</td>
<td>0.074</td>
<td>−0.026</td>
<td>0.0%</td>
</tr>
<tr>
<td>MFGorb</td>
<td>L</td>
<td>−0.215</td>
<td>0.252</td>
<td>0.5%</td>
</tr>
<tr>
<td>MidPar</td>
<td>L</td>
<td>0.116</td>
<td>−0.051</td>
<td>0.0%</td>
</tr>
<tr>
<td>mFC</td>
<td>L</td>
<td>−0.125</td>
<td>0.257</td>
<td>0.0%</td>
</tr>
<tr>
<td>PostPar</td>
<td>L</td>
<td>0.083</td>
<td>0.096</td>
<td>0.0%</td>
</tr>
<tr>
<td>PresG</td>
<td>L</td>
<td>0.078</td>
<td>0.048</td>
<td>0.0%</td>
</tr>
<tr>
<td>SFG</td>
<td>L</td>
<td>0.180</td>
<td>0.025</td>
<td>0.0%</td>
</tr>
<tr>
<td>AntPar</td>
<td>R</td>
<td>0.016</td>
<td>−0.077</td>
<td>0.0%</td>
</tr>
<tr>
<td>IFGpop</td>
<td>R</td>
<td>−0.011</td>
<td>0.075</td>
<td>0.0%</td>
</tr>
<tr>
<td>Insula</td>
<td>R</td>
<td>−0.185</td>
<td>0.227</td>
<td>0.0%</td>
</tr>
<tr>
<td>MFG</td>
<td>R</td>
<td>0.058</td>
<td>−0.006</td>
<td>0.0%</td>
</tr>
<tr>
<td>MFGorb</td>
<td>R</td>
<td>−0.004</td>
<td>0.019</td>
<td>0.0%</td>
</tr>
<tr>
<td>MidPar</td>
<td>R</td>
<td>0.040</td>
<td>−0.110</td>
<td>0.0%</td>
</tr>
<tr>
<td>mFC</td>
<td>R</td>
<td>−0.321</td>
<td>0.440</td>
<td>0.0%</td>
</tr>
<tr>
<td>PostPar</td>
<td>R</td>
<td>−0.312</td>
<td>0.434</td>
<td>0.0%</td>
</tr>
<tr>
<td>PresG</td>
<td>R</td>
<td>0.034</td>
<td>0.118</td>
<td>0.0%</td>
</tr>
<tr>
<td>SFG</td>
<td>R</td>
<td>0.066</td>
<td>−0.034</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

responses against responses from the same brain region in all other
participants exposed to that stimulus. This ceiling is designed to quantify
the variance that can be explained based on the stimuli alone, inde-
pendently of inter-participant variation. As shown, models explain a
substantial amount of the available variance in LANG. MD models
explain no variance on the evaluation set, suggesting that the MD model
did not learn generalizable patterns.

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Because fROIs were modeled as random effects in these analyses,
pairwise statistical testing of between-region differences in effect
amplitude is not straightforward, and systematic investigation of re-

Table 5
Model effect estimates.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>LANG</th>
<th>MD</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>% Total</td>
<td>% Relative</td>
<td>% Held-Out Variance Explained</td>
</tr>
<tr>
<td>Sound Power</td>
<td>−0.055</td>
<td>−0.006</td>
<td>0.0%</td>
<td>−0.003</td>
</tr>
<tr>
<td>TR Number</td>
<td>−0.148</td>
<td>0.048</td>
<td>0.0%</td>
<td>−0.005</td>
</tr>
<tr>
<td>Rate</td>
<td>0.242</td>
<td>0.146</td>
<td>0.048</td>
<td>0.004</td>
</tr>
<tr>
<td>Frequency</td>
<td>−0.060</td>
<td>−0.199</td>
<td>0.134</td>
<td>0.003</td>
</tr>
<tr>
<td>5-gram Surprisal</td>
<td>0.209</td>
<td>−0.025</td>
<td>0.003</td>
<td>0.038</td>
</tr>
<tr>
<td>PCFG Surprisal</td>
<td>0.235</td>
<td>0.097</td>
<td>0.003</td>
<td>0.576</td>
</tr>
<tr>
<td>Network</td>
<td>–</td>
<td>–</td>
<td>−1.32</td>
<td></td>
</tr>
<tr>
<td>Sound Power by Network</td>
<td>–</td>
<td>–</td>
<td>−0.050</td>
<td></td>
</tr>
<tr>
<td>TR Number by Network</td>
<td>–</td>
<td>–</td>
<td>−0.008</td>
<td></td>
</tr>
<tr>
<td>Rate by Network</td>
<td>–</td>
<td>–</td>
<td>0.269</td>
<td></td>
</tr>
<tr>
<td>Frequency by Network</td>
<td>–</td>
<td>–</td>
<td>0.040</td>
<td></td>
</tr>
<tr>
<td>5-gram Surprisal by Network</td>
<td>–</td>
<td>–</td>
<td>0.212</td>
<td></td>
</tr>
<tr>
<td>PCFG Surprisal by Network</td>
<td>–</td>
<td>–</td>
<td>0.193</td>
<td></td>
</tr>
</tbody>
</table>

Table 6
Model percent variance explained compared to a “ceiling” linear model regressing against the mean response of all other participants for a particular story/fROI. “% Total” columns show absolute percent variance explained, while “% Relative” columns show percent variance explained relative to the ceiling.

<table>
<thead>
<tr>
<th></th>
<th>LANG</th>
<th>MD</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Total</td>
<td>% Relative</td>
<td>% Total</td>
</tr>
<tr>
<td>Ceiling</td>
<td>6.18%</td>
<td>100%</td>
<td>1.34%</td>
</tr>
<tr>
<td>Model (train)</td>
<td>3.68%</td>
<td>59.5%</td>
<td>0.75%</td>
</tr>
<tr>
<td>Model (evaluation)</td>
<td>2.30%</td>
<td>37.2%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Table 7
LANG result. Significance in LANG by paired permutation test of log-likelihood improvement on the evaluation set from including a fixed effect for each of 5-gram Surprisal and PCFG Surprisal, over (1) a baseline with neither fixed effect and (2) baselines containing the other fixed effect only. The Effect Estimate column shows the estimated effect size from the model containing the fixed effect (i.e. the area under the estimated HRF).

<table>
<thead>
<tr>
<th>Comparison</th>
<th>p</th>
<th>LL Improvement</th>
<th>Effect Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-gram over neither</td>
<td>0.0001 ***</td>
<td>182</td>
<td>0.307</td>
</tr>
<tr>
<td>PCFG over neither</td>
<td>0.0001 ***</td>
<td>183</td>
<td>0.352</td>
</tr>
<tr>
<td>5-gram over PCFG</td>
<td>0.0001 ***</td>
<td>61</td>
<td>0.209</td>
</tr>
<tr>
<td>PCFG over 5-gram</td>
<td>0.0001 ***</td>
<td>61</td>
<td>0.235</td>
</tr>
</tbody>
</table>
processing load elsewhere in the brain through hemodynamic factors ("vascular steal") (Lee et al., 1995; Saad et al., 2001; Harel et al., 2002; Kannurpatti and Biswal, 2004) and/or neuronal ones such as inhibition by an attention mechanism (Smith et al., 2000; Shmuel et al., 2002, 2006). The ways by which such mechanisms might give rise to negative frequency effects in our data are not currently clear. Since frequency effects are not central to our present research question, we leave targeted investigation of their existence and direction to future research.

Fig. 5 and Table 10 assess the generalizability of surprisal effects across participants. Fig. 5 shows most-by-participant improvements clustered around a positive median, without strong visual indication of large-magnitude positive outliers that might exclusively drive the effect. This intuition is quantified in Table 10. As shown, held-out likelihood improves for most participants in every comparison, and at least 5 of the most responsive participants can be removed in each comparison without changing the significance of the effect.

### Table 8
**MD result.** Significance in MD by paired permutation test of log-likelihood improvement on the evaluation set from including a fixed effect for each of 5-gram Surprisal and PCFG Surprisal, over (1) a baseline with neither fixed effect and (2) baselines containing the other fixed effect only. A p-value of 1.0 is assigned by default to comparisons in which held-out likelihood improved under ablation. The *Effect Estimate* column shows the estimated effect size from the model containing the fixed effect (i.e. the area under the estimated HRF).

<table>
<thead>
<tr>
<th>Comparison</th>
<th>p</th>
<th>LL Improvement</th>
<th>Effect Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-gram over neither</td>
<td>0.137</td>
<td>3</td>
<td>0.019</td>
</tr>
<tr>
<td>PCFG over neither</td>
<td>1.0</td>
<td>−29</td>
<td>0.081</td>
</tr>
<tr>
<td>5-gram over PCFG</td>
<td>1.0</td>
<td>−8</td>
<td>−0.025</td>
</tr>
<tr>
<td>PCFG over 5-gram</td>
<td>1.0</td>
<td>−40</td>
<td>0.097</td>
</tr>
</tbody>
</table>

### Table 9
**Combined result.** Significance in the combined data by paired permutation test of log-likelihood improvement on the evaluation set from including a fixed interaction for each of 5-gram Surprisal and PCFG Surprisal with Network, over (1) a baseline with neither fixed interaction and (2) baselines containing the other fixed interaction only. The *Effect Estimate* column shows the estimated interaction size from the model containing the fixed interaction (i.e. the difference in effect estimate between LANG and MD).

<table>
<thead>
<tr>
<th>Comparison</th>
<th>p</th>
<th>LL Improvement</th>
<th>Effect Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-gram:Network over neither</td>
<td>0.00001***</td>
<td>144</td>
<td>0.212</td>
</tr>
<tr>
<td>PCFG:Network over neither</td>
<td>0.00001***</td>
<td>144</td>
<td>0.193</td>
</tr>
<tr>
<td>5-gram:Network over PCFG</td>
<td>0.00001***</td>
<td>53</td>
<td>0.301</td>
</tr>
<tr>
<td>PCFG:Network over 5-gram</td>
<td>0.00001***</td>
<td>53</td>
<td>0.317</td>
</tr>
</tbody>
</table>

### Table 10
**Generality of LANG surprisal effects across participants.** Median likelihood improvement in LANG on the evaluation set by participant, percent of participants whose held-out predictions improved due to surprisal effects, and the number of participants with the largest held-out improvement whose data can be removed without changing the significance of the effect at a 0.05 level. Held-out likelihood improves for most participants in every comparison, and at least 5 of the most responsive participants can be removed in each comparison without changing the significance of the effect.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Median LL Improvement by Participant</th>
<th>% Participants Improved</th>
<th>Num Removable Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-gram over neither</td>
<td>1.236</td>
<td>71.8%</td>
<td>19</td>
</tr>
<tr>
<td>PCFG over neither</td>
<td>0.732</td>
<td>64.1%</td>
<td>14</td>
</tr>
<tr>
<td>5-gram over PCFG</td>
<td>0.335</td>
<td>61.5%</td>
<td>7</td>
</tr>
<tr>
<td>PCFG over 5-gram</td>
<td>0.498</td>
<td>60.3%</td>
<td>5</td>
</tr>
</tbody>
</table>

6. **Discussion**

The current study examined signatures of predictive processing during naturalistic story comprehension in two functionally distinct cortical networks: the domain-specific language (LANG) network, and the domain-general multiple demand (MD) network. Specifically, we tested which of these networks increased their responses with lower word predictability, operationalized using both 5-gram and probabilistic context-free grammar (PCFG) surprisal. The main results, yielded by continuous-time deconvolutional regression (CDR) analysis of surprisal effects in the two networks, are shown in Tables 7–9: in LANG, both 5-gram Surprisal and PCFG Surprisal have positive effects that yield statistically significant improvements to held-out likelihood, both over a baseline containing neither fixed effect as well as over one another. By contrast, in MD, neither surprisal effect is significant in any comparison. A direct test for a difference in surprisal effects across the two networks (Table 9) shows that the interactions of both surprisals with network are positive and statistically significant, indicating that the BOLD response to both surface-level (5-gram) and structural (PCFG) word predictability is larger in LANG than MD. These results are over a baseline that includes an effect for lexical frequency (log unigram probability), which is notable given the strong natural correlation between surprisal and frequency, both generally (Dembarg and Keller, 2008) and in the current experimental materials (r = 0.78 overall). This finding suggests that the surprisal effects reported here are indeed driven by predictive coding and not merely by the cost of retrieving infrequent words. Together, these results demonstrate that predictive coding for upcoming words is primarily a canonical computation carried out by domain-specific cortical circuits, rather than by feedback from higher, domain-general executive control circuits, and that these predictions depend on both surface-level and structural information sources. Our finding of a generalized effect of PCFG Surprisal throughout the language network aligns with prior findings of evidence for linguistic prediction (e.g. Kuperberg et al., 2000; Baumgaertner et al., 2002; Friederici et al., 2005; Obleser et al., 2007) and syntactic processing (e.g., Blank et al., 2016; see Zaccarella et al., 2017 for review) in these regions, but suggests that prior evidence of linguistic prediction effects in MD (e.g. Kuperberg...
null
memory-based estimators of structural processing have been studied across behavioral and cognitive neuroscience investigations, including embedding difference (Wu et al., 2010), the number of open nodes based on a particular parsing strategy (top-down, bottom-up, or left-corner; Nelson et al., 2017; Brennan & Pyllkänen, 2017), dependency locality costs (storage or integration cost from maintaining and retrieving syntactic dependencies; Gibson, 2000), and encoding or retrieval interference (i.e. processing costs in the ACT-R framework; Lewis and Vaisnīth, 2005). Effort due to memory storage and retrieval is plausibly distinct from effort due to reallocating resources between competing structural interpretations of the unfolding sentence (a standard interpretation of surprisal effects, e.g. Hale, 2001; Levy, 2008) (but see Futrell & Levy, 2017), and a complete account of human language processing will likely involve both prediction-based and integration-based computations (Levy et al., 2013; Levy & Gibson, 2013). Although these kinds of integration effects are outside the scope of the present study of predictive coding, we note that some have argued that prediction may subserve memory retrieval and therefore interact with integrative processing (Altman, 1998). Therefore, the prediction effects reported here may, to some extent, be amenable to interpretation as effects of integration. That is, researchers who view “prediction” as a conscious lexically specific activity may view these results as evidence of conceptual pre-activation or preparedness that eases integration once a word is observed (see Ferreira and Chantavarin, 2018, for an overview of this distinction). We leave to future work a fuller investigation of this distinction and simply note that our results indicate that any such pre-activation processes appear to be restricted to the LANG network, rather than invoking the MD network, and are strongly correlated with probabilistic measures of word predictability. The fMRI dataset produced by this study will hopefully support further investigation into the interplay of memory and expectation in the language-selective and domain-general networks.

Our emphasis on structural influences on prediction, rather than sensitivity to syntactic structure more generally, is a possible explanation for one apparent discrepancy between our results and those of some previous studies. In particular, we do not find clear evidence of PCFG Surprisal effects in the AntTemp language fROI (relative to other effect sizes in this study, the PCFG Surprisal estimate in AntTemp – 0.017 – is close to zero), whereas numerous previous studies have argued for syntactic effects in left anterior temporal cortex, both using hand-constructed stimuli (Mazoyer et al., 1993; Stowe et al., 1998; Frieder close to zero), whereas numerous previous studies have argued for syntactic effects in the AntTemp – 0.017 – is Surprisal effects in the AntTemp language fROI (relative to other effect
domain-general networks.

In summary, our findings based on a large-scale naturalistic fMRI experiment support a view of linguistic prediction as implemented by domain-specific cortical circuits, sensitive to both surface-level and syntactic information sources, and generalized across the functional language network.

Declaration of competing interest

The authors declare no competing financial interests.

CRediT authorship contribution statement

Cory Shain: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing. Idan Asher Blank: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing - review & editing. William Schuler: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Writing - review & editing. Evelina Fedorenko: Conceptualization, Formal analysis, Funding acquisition, Writing - review & editing.

Acknowledgments

This research was supported by NIH award R00-HD-057522 (E.F.) and by National Science Foundation grant #1816891 (W.S.). All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation. E.F. was additionally supported by NIH awards R01-DC-016607 and R01-DC-016950 and by a grant from the Simons Foundation via the Simons Center for the Social Brain at MIT. The authors would also like to acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology (MIT), and the support team (especially Steven Shannon and Atsushi Takahashi). The authors also thank Nancy Kanwisher and Ted Gibson for recording the stories, and EvLab members for help with data collection, especially Zach Mineroff and Alex Paunov.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2019.107307.

References


Amunts, K., Dehaene, S., 2019. Combination of functional and structural analysis reveals neural basis of linguistic predictions during sentence and fMRI reveals neural basis of linguistic predictions during sentence.


D’Antonio, L., Rief, W., 2017. Learning about expectation violation from prediction error paradigms_A meta-analysis on brain processes following a prediction error. Front. Psychol. 8, 1253.


