Neuropsychologia 146 (2020) 107479

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: http://www.elsevier.com/locate/neuropsychologia

Localizing syntactic predictions using recurrent neural network grammars

Jonathan R. Brennan [a,*], Chris Dyer [b], Adhiguna Kuncoro [c], John T. Hale [d]

[a] University of Michigan, USA
[b] DeepMind, London, UK
[c] DeepMind, Oxford University, Oxford, UK
[d] DeepMind, University of Georgia, USA

ARTICLE INFO

Keywords:
Syntax
Parsing
Deep learning
Language model
Surprisal
fMRI

ABSTRACT

Brain activity in numerous perisylvian brain regions is modulated by the expectedness of linguistic stimuli. We leverage recent advances in computational parsing models to test what representations guide the processes reflected by this activity. Recurrent Neural Network Grammars (RNNGs) are generative models of (tree, string) pairs that use neural networks to drive derivational choices. Parsing with them yields a variety of incremental complexity metrics that we evaluate against a publicly available fMRI data-set recorded while participants simply listen to an audiobook story. Surprisal, which captures a word's un-expectedness, correlates with a wide range of temporal and frontal regions when it is calculated based on word-sequence information using a top-performing LSTM neural network language model. The explicit encoding of hierarchy afforded by the RNNG additionally captures activity in left posterior temporal areas. A separate metric tracking the number of derivational steps taken between words correlates with activity in the left temporal lobe and inferior frontal gyrus. This pattern of results narrows down the kinds of linguistic representations at play during predictive processing across the brain's language network.

1. Introduction

This study concerns the kind of information that is used to shape expectations during language comprehension. Predictive processing has long been seen as central to rapid and efficient language comprehension (Tanenhaus et al., 1995; Marslen-Wilson, 1975). One mechanism by which prediction serves rapid comprehension is by pre-computing or pre-activating linguistic representations based on contextual cues and other top-down information. Preactivation is an important principle across different levels of comprehension, including lexical processing (Kutas and Federmeier, 2000) and syntactic parsing (Hale, 2014). On a neural level, prior work has found evidence that top-down information can propogate through, or cascade, from higher-level conceptual representations to lower-level lexical and perhaps perceptual stages (Dikker and Pylkkänen, 2012; Molinaro et al., 2013; but c.f. Nieuwland et al., 2018). The notion that top-down predictions can simultaneously impact multiple stages or levels of processing is consistent with neuroimaging research showing that the expectancy of a word impacts neural activation in a broad range of language-related brain regions, including the temporal lobes bilaterally and inferior frontal regions (Willems et al., 2016; Lopopolo et al., 2017; Henderson et al., 2016; Lowder et al., 2018; Brennan et al., 2016). We test here whether activity in these regions is conditioned by similar, or by different, aspects of linguistic context, with a specific focus on syntactic structure.

Our study builds on recent work that has probed how neural activity reflects predictive processing using the linking hypothesis of "surprisal." This quantity, which comes from information theory, measures the conditional probability of a word given some characterization of its context (see Hale, 2016 for an introduction and review with a focus on psycholinguistics.) At the sentence level, surprisal has been found to correlate with reading times (Hale, 2001), eye-tracking measures (Boston et al., 2008; Demberg and Keller, 2008; Roark et al., 2009; Frank and Bod, 2011), electrophysiological responses (Frank et al., 2015; Hale et al., 2018; Brennan and Hale, 2019; Nelson et al., 2017; Brennan et al., 2018), and functional magnetic resonance imaging signals (fMRI; Willems et al., 2016; Lopopolo et al., 2017; Henderson et al., 2016; Lowder et al., 2018; Brennan et al., 2016).

Focusing on spatially localized results from fMRI, brain regions whose activity is moderated by surprisal show overlap, but also interesting differences, across several studies. For example, Lopopolo et al. (2017) report activations in the left and right temporal lobes, both anterior and posterior overlapping the middle and superior temporal

* Corresponding author.
E-mail address: jobrenn@umich.edu (J.R. Brennan).

https://doi.org/10.1016/j.neuropsychologia.2020.107479
Received 18 March 2019; Received in revised form 3 April 2020; Accepted 29 April 2020
Available online 16 May 2020
0028-3932/© 2020 Elsevier Ltd. All rights reserved.

gyri, and they also see activity in the superior frontal gyrus. Lopopolo et al. define surprisal for each word's part-of-speech using a context comprising the two immediately preceding words (a third-order Markov language model.) Crucially, surprisal values based on different language models yield different results. Willems et al. (2016) use a similar Markov language model, but do so for each word individually (not for parts-of-speech). They report activation for surprisal that is constrained principally to the temporal lobes bilaterally. In contrast, Henderson et al. (2016) compute surprisal based on a context that explicitly encodes phrase structure but, like Lopopolo et al. (2017), they examine parts-of-speech. They see effects primarily in the left inferior frontal gyrus and also the left anterior temporal lobe.

Brennan et al. (2016), who also use fMRI and the surprisal linking hypothesis, apply statistical model comparison to tease apart expectation effects as a function of whether they might reflect more sequence-based information, as in the Markov model, or hierarchical phrase-structure information. In this effort, they build on prior work with behavioral and electrophysiological measures by Frank and Bod (2011) and Frank et al. (2015) who themselves are entering into a larger debate in the language sciences about the role of abstract grammatical structure in online language processes (for discussion, see e.g. Caplan, 1992, ch. 7, and Lewis and Phillips, 2015.) Focusing on expectations about a word's part-of-speech, Brennan et al. report that sequence-based Markov models correlate with activity in the anterior and posterior temporal lobe, as well as the left inferior frontal gyrus. They also find effects when surprisal is defined in terms of a hierarchical phrase-structure grammar in those regions as well as a temporal-parietal region and a left premotor region. And, further, the effects for hierarchical structure are found above-and-beyond effects for the sequence-based models in all regions examined except the left inferior frontal gyrus. The region of interest data used for those results are available in the public domain, and they form the dataset that we analyze in this study.

To better understand the role of sequence-based and hierarchical information in guiding neural mechanisms for prediction, the present study draws on computational modeling from Hale et al. (2018) to extend in two specific ways our understanding of the localization of surprisal effects. First, we go beyond Markov models by using state-of-the-art Long Short-Term Memory or LSTM recurrent neural networks (e.g. Hochreiter and Schmidhuber, 1997; Mikolov et al., 2010). Such networks operate over word sequences and do not explicitly encode phrase structure. But, rather than constraining prior context to a fixed window, as in a Markov model, or using a simple recurrent network with a strong proximity bias (as in electrophysiological work like Frank et al., 2015 and Brennan and Hale, 2019), these networks function to dynamically up-weight or down-weight different aspects of context, encoded in hidden layers. In this way, they may implicitly recover structural relations when demanded by the utility of that information as determined during training (Linzen et al., 2016; Gulordava et al., 2018; Wilcox et al., 2019). They are thus a powerful baseline for comparing against models that explicitly encode hierarchical information.

Second, we adopt a hierarchical parsing model that accommodates the syntactic ambiguity common in natural language. While ambiguity is a frequent locus of study in psycholinguistics, prior work on hierarchical sentence structure, as seen through the lense of surprisal, has relied on one-path "gold-standard" syntactic parses to determine the syntactic context. To capture ambiguity, we adopt a parser based on Recurrent Neural Network Grammars (RNNG; Dyer et al., 2016) introduced by Hale et al. (2018). This parser achieves competitive performance on standard natural language processing metrics and also, by virtue of its account of ambiguity and its capacity to explicitly encode abstract hierarchical sentence structure, offers a compelling tool to estimate cognitive correlates of hierarchical processing. Using electroencephalography (EEG) data collected while participants listen to an audiobook, Hale et al. report responses that are modulated by surprisal and that are specific to the hierarchical component of their parser, above-and-beyond sequence-based surprisal effects estimated from a LSTM model.

Here, we use the same computational models as Hale et al. but now apply them to spatially precise fMRI data. Doing so allows us to address which kinds of syntactic information impact processing that is distributed across the set of language-related frontal and temporal regions that were described above.

In sum, to better understand how sequence-based and hierarchical information might be used to guide expectations across language-related brain regions, we combine neural network-based language models that either do or do not explicitly encode hierarchy with an openly available fMRI dataset collected while participants listened to an audiobook. To preview our results: we find broad effects across regions for surprisal, but the influence of explicit hierarchy on surprisal, derived from the recurrent neural network grammar, is constrained to posterior areas of the left temporal lobe.

2. Methods

2.1. Stimulus & fMRI data

We analyze fMRI time-series from 26 individuals in six regions of interest (ROI). These come from 28 publicly available datasets[1] that were recorded while participants listened to a 12.4 min audiobook story (the first chapter of Alice's Adventures in Wonderland) (Brennan et al., 2016). We set aside two of the available datasets because those participants scored at chance on a post-scan comprehension questionnaire. Equipment and scan protocol details are available in Brennan et al. (2016).

The time-series come from six ROIs that were localized in each individual using a combination of functional and anatomical criteria that we summarize here. First, a statistical analysis identified voxels whose activity increased linearly with the rate that words appear in the story. This is a "broad brush" localizer that should reveal activation for regions involved in many different aspects of language comprehension spanning acoustic, phonological, lexical, syntactic, and semantic analysis. Peaks in this functional localizer exceeding at least $t = 2$ were identified in six anatomically-defined regions: the left anterior temporal lobe (LATL), the right anterior temporal lobe (RATL), the left inferior frontal gyrus (LIFG), the left posterior temporal lobe (LPTL), the left inferior parietal lobule (LIPL), and a left posterior middle frontal gyrus premotor region (LPreM). The intersection of functional peaks with anatomical regions was based on the Harvard-Oxford probabalistic brain atlas. Not all participants showed a functional peak in each anatomical region (Exclusions: LATL = 3, RATL = 4, LIFG = 5, LPTL = 2, LIPL = 2, LPreM = 2). For each ROI, the fMRI signal was averaged within a 10 mm radius sphere around the functional peak.

The data thus constitute, for each participant, up to six time-series each of 362 samples (12.4 min = 744 s = 372 samples at TR = 2; the first 10 samples of data were discarded.) The individually-defined ROI peaks are shown in Fig. 2A.

2.2. Computational modeling

These fMRI time-series are modeled using metrics derived from computational models of sentence processing presented by Hale et al. (2018). To model the structural aspect of human language comprehension, we use Recurrent Neural Network Grammars (RNNG) (Dyer et al., 2016; Kuncoro et al., 2016).

[1] The data are available for download at https://sites.lsa.umich.edu/cnllab/2016/06/11/data-sharing-fmri-timecourses-story-listening/.

2.2.1. Example-based introduction to RNNGs

Putting RNNGs into an incremental parsing system yields a program that takes as input English words one by one in the order they would be spoken or heard in time. The output of this program is a phrase structure tree whose labels are consistent with the Penn Treebank (Marcus et al., 1993). To build these trees, the RNNG-based incremental parser stochastically chooses a sequence of actions such as the ones shown in Table 1. The only options are (a) to open a new phrase (i.e. posit a new tree node) (b) to close-off a phrase (akin to the "reduce" action in a bottom-up shift-reduce algorithm) or (c) move on to the next word ("generate"). Actions are selected on the basis of a learned numerical vector, symbolized by the term s_t in Fig. 1.

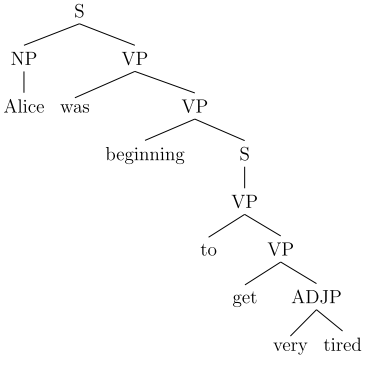
This *Syntactic Context* vector is itself determined by a numerical encoding of the current contents of a stack memory (Dyer et al., 2015); this is schematized in Fig. 1A. This "stack LSTM" allows classic symbolic parsing techniques (see chapter 3 of Hale, 2014, inter alia) to become deep neural networks. These networks are trained in a supervised manner on action sequences that correspond uniquely to already-parsed sentences that are provided as training data. During training, when the neural network guesses the wrong parser action, error is back-propagated through the entire system. This can affect many different trainable parameters, including *Syntactic Context* itself. While vector representations like *Syntactic Context* are not directly interpretable, one of the things *Syntactic Context* is likely representing is the history of previous parser actions. If this is true, then the architecture as a whole qualifies as a relaxation of restrictive context-free assumptions that are inherent in more classical approaches to phrase structure parsing — such as probabilistic context-free grammars. At the same time, in virtue of having categorical representations that can be interpreted as phrase structure trees it can potentially make use of tree-structural relationships such as c-command in deciding which action to take.

We describe the elements of this parser in general terms here, with a focus on those components that bear on interpreting the fMRI data. For greater detail, see Dyer et al. (2016); Kuncoro et al. (2016); Hale et al. (2018).

2.2.2. Key aspects of the parsing model

One such element is the stack memory shown in Fig. 1A. Classical symbolic parsing systems would employ such a stack to keep track of partially-built phrase structure at a given point during incremental processing. For instance, it might contain the depth-1 tree whose root label is NP and whose three daughters are "the" "rabbit" and "hole." Whenever the RNNG-based parser opts to close off a phrase, such as this NP, a composition function is triggered in which the elements on the stack that comprise that phrase and the phrasal label itself are popped off and encoded into a single element. Inclusion of this function means that phrasal composition is explicitly encoded within the RNNG. Following Hale et al. (2018), we also explore an alternative architecture which does not include this composition function. The alternative, dubbed RNNG$_{-COMP}$, places on the stack a sequence of bracketed expressions but does not explicitly encode whether those expressions belong to a single (composed) term. Specifically, RNNG$_{-COMP}$ replaces the "close phrase" action (EndPhrase in Fig. 1B) with a different action that adds a special symbol to the stack indicating the end of a phrase. For example, the first row of Table 2 shows stack contents from the RNNG that is illustrated in Fig. 1. For RNNG$_{-COMP}$, the same information would instead be encoded by a sequence of seven symbols on the stack as illustrated in row 2 of Table 2. This table shows how, in RNNG$_{-COMP}$, the individual daughters of NP are each held on the stack separately, rather than being composed into a single representation. Without the composition mechanism, it becomes necessary to employ special right-bracket symbols like $)_{NP}$ in order to be able to correctly build trees.

Table 1
An example of a sequence of RNNG parser actions used in the analysis of Alice's Adventures in Wonderland.

open phrase S	
open phrase NP	
generate word *alice*	
close off phrase	
open phrase VP	
generate word *was*	
open phrase VP	
generate word *beginning*	
open phrase S	
open phrase VP	
generate word *to*	
open phrase VP	
generate word *get*	
open phrase ADJP	
generate word *very*	
generate word *tired*	

As is the case in human sentence comprehension, the RNNG parser does not have the capacity to look ahead to upcoming words. Thus, decisions made based on a partial word sequence may turn out to be incorrect based on subsequent words. We use the standard technique of "beam search" to navigate this non-determinism (Roark, 2004). The parser keeps track of a set of candidate analyses, the "beam", which are ranked based on the probabilities of their constituent actions. This approach matches the broad adoption of ranked-parallel parsing in psycholinguistics (e.g. Gibson, 1991; Jurafsky, 1996). In the case of a generative model like the RNNG, the beam search algorithm must take into account the imbalance in probability between structural actions and lexical actions (Stern et al., 2017). The "word-synchronous" search algorithm from Stern et al. resolves this imbalance by searching through structural actions until a sufficent number of candidate analyses in the beam take a lexical action; see Algorithm 1 from Hale et al. (2018), repeated in the Supplemental Materials, for the implementation used here. The number of analyses needed to achieve a synchronous state is a free parameter k in this model; we set $k = 200$ as this value yielded a good fit to neural signals recorded using electroencephalography (EEG) in prior work (Hale et al., 2018). The number of analyses in the beam that are carried forward from each word is fixed at $k/10 = 20$.

As mentioned earlier the network weights of the RNNG, as well as those in the various baseline models discussed below, are all set via Backpropagation Through Time on a common training corpus made up of chapters 2–12 of Alice's Adventures in Wonderland, comprising 24,941 words. The corpus was parsed using the Stanford parser (Klein and Manning, 2003) yielding a syntactic annotation in accordance with the Penn Treebank annotation scheme (Marcus et al., 1993). Further training details are given in Hale et al. (2018).

While our principle goal is to quantify model quality in terms of how well they fit human neural signals, we can also quantify the quality of the model in terms of the text itself. *Perplexity* is a common metric which summarizes the average uncertainty of a language model when confronted with text (Jelinek, 1998), computed as:

$$2^{-\frac{1}{N}\sum_{w=1}^{N} log_2(Pr(w|\mathrm{Ctxt}))}$$

where $Pr(w|\mathrm{Ctxt})$ is the model-estimated probability of word w given its *Syntactic Context*. When computed for the first chapter of Alice's Adventures in Wonderland, our models show values of 25.01, 23.16, and 24.10, for the full RNNG, RNNG$_{-COMP}$, and a LSTM baseline (described below), respectively.[2] These are lower than those typically observed in computational linguistics, which likely reflects the relatively small genre-specific training text used. For present purposes, what is

[2] Note that RNNG perplexity reflects the joint probability of (tree, string) pairs while LSTM only reflects string probabilities.

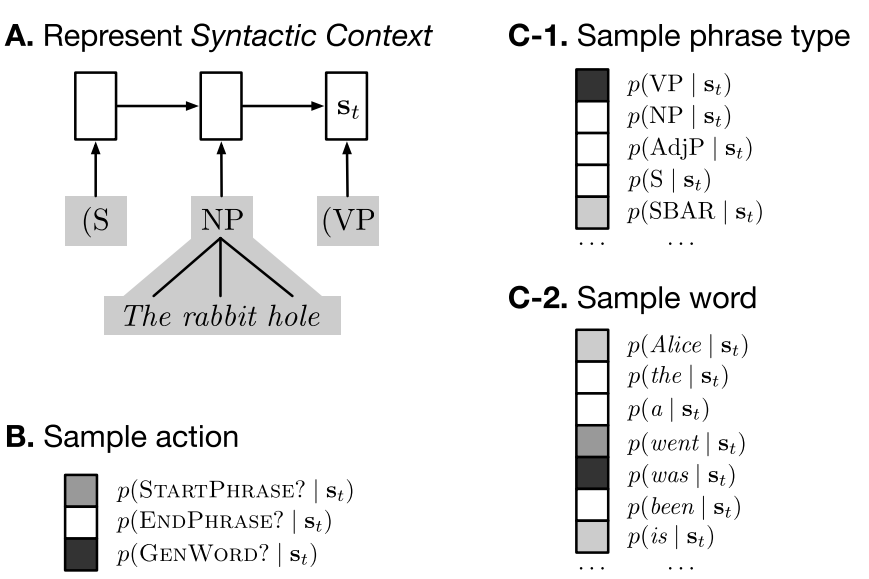


Fig. 1. Data-flow diagram for the RNNG used in this paper. The arrows indicate the order in which the RNNG updates a vector representation of the *Syntactic Context* as it processes an input sequence. The final such state is s_t. (A) In this example, the *Syntactic Context* s_t includes, among other pieces, a composed Noun Phrase with daughters *the rabbit hole*. (B) The parser makes a decision about which of three actions to take based on s_t. Depending on the action taken, the parser stochastically generates (C-1) phrasal labels or (C-2) individual lexical items.

important is that these values are much smaller than would be expected if the models had not learned any regularities in the text (under a uniform language model, perplexity is equal to vocabulary size and there are about 2,900 word types in the present case.) And, further, these show that the models all perform roughly equally in terms of predicting the text itself.

The next sections turn to how we use these models to derive predictions for the fMRI data collected from participants listening to the audiobook.

2.2.3. *Complexity metrics used in this study*

To link the RNNG parsing model with neural fMRI data, we define two *complexity metrics* that quantify different aspects of the model's state at each word. As such, these metrics tap into different aspects of processing complexity. The principle aim of this paper is to probe predictive syntactic processing. To this end we use the information theoretic quantity of SURPRISAL to model the (un)expectedness of a word given its syntactic left-context (Hale, 2001, 2016; Boston et al., 2008); SURPRISAL is calculated as the log-ratio of the forward probabilities of the analyses in the beam:

$$\log_2\left(\frac{\text{BeamAfter}}{\text{BeamBefore}}\right)$$

Or, equivalently (see Hale, 2016):

$$-\log_2(\Pr(w_i|\text{Ctxt}_{w_0 \ldots w_{i-1}}))$$

This metric is higher when encountering a word that is unexpected, and lower when a word is expected. The word-by-word nature of SURPRISAL distinguishes it from perplexity, defined above, which rates the average uncertainty of a language model on a text. Prior work has shown a good match between SURPRISAL and, for example, the N400 event-related potential (ERP) brain response associated with lexical processing (Frank et al., 2015). Important for our purposes, SURPRISAL values vary depending on how the (left-)context for forward probabilities is defined. ERP research has shown that contexts that make hierarchical information explicit, as with RNNGs, yield SURPRISAL values that show better fits to ERP signals than when SURPRISAL is calculated using sequence information alone (Hale et al., 2018; Brennan and Hale, 2019).

As in Hale et al. (2018), we define several different models for computing SURPRISAL. In addition to the full RNNG, we also compute surprisal when the explicit composition mechanism is removed (RNNG$_{-COMP}$). We further compute SURPRISAL from a state-of-the-art LSTM recurrent neural network (e.g. Mikolov et al., 2010). Rather than explicitly encoding phrase structure, such networks operate over word sequence information although they may implicitly recover structural relations through training (Linzen et al., 2016; Gulordava et al., 2018). This key difference between LSTM and the RNNG models is illustrated in row 3 of Table 2 which contrasts the sequence information available to the LSTM with the structural *Syntactic Context* available to the RNNG models in the rows above.[3] As already mentioned in the Introduction, we predict that SURPRISAL should affect a broad range of stages of language comprehension and, thus, we expect this measure to modulate activity across the language-related ROIs we probe. The key theoretical question is whether, and in which regions, SURPRISAL values based on explicit hierarchical structure outperform those values based on the sequence-based LSTM baseline.

The RNNG model also affords another complexity metric which seeks to capture the intuitive notion of sentence processing "work." This metric is called DISTANCE and it can be understood as a more realistic version of the node-counting metrics used in Brennan et al. (2012, 2016). The additional realism comes from explicitly modeling the ambiguity resolution process, which node-counting metrics do not address. These classical measures only tabulate effort spent building the correct tree, without regard to effort spent sifting through alternative trees. This latter factor would be required in an incremental parsing system that actually copes with ambiguity. DISTANCE models this factor, over and above the basic equation between tree nodes and parser effort. It counts the number of syntactic analyses across the entire beam considered by the parser in terms of individual actions as illustrated in Table 1. This amounts to a penalty for cases where the RNNG-based incremental parser has a hard time finding phrase structures that span the next input word. DISTANCE is roughly analogous to the "Arcs Attempted" metric in Kaplan (1972).

As it is defined in terms of the search procedure rather than the grammar, DISTANCE does not map on to any particular syntactic relationship (e.g. it is not c-command). Nor does it map on to expectedness of a word in the way that SURPRISAL does. Rather the DISTANCE metric connects best with efforts to probe the neural bases of combinatoric operations themselves (e.g. Pylkkänen, 2016; Brennan and Pylkkänen,

[3] LSTM language models are a powerful baseline for comparing against the RNNG as the former is capable of recovering structural dependencies, but only does so as demanded by the utility of that information as determined during training. This contrasts with the RNNG, for which the architecture forces a structural bias.

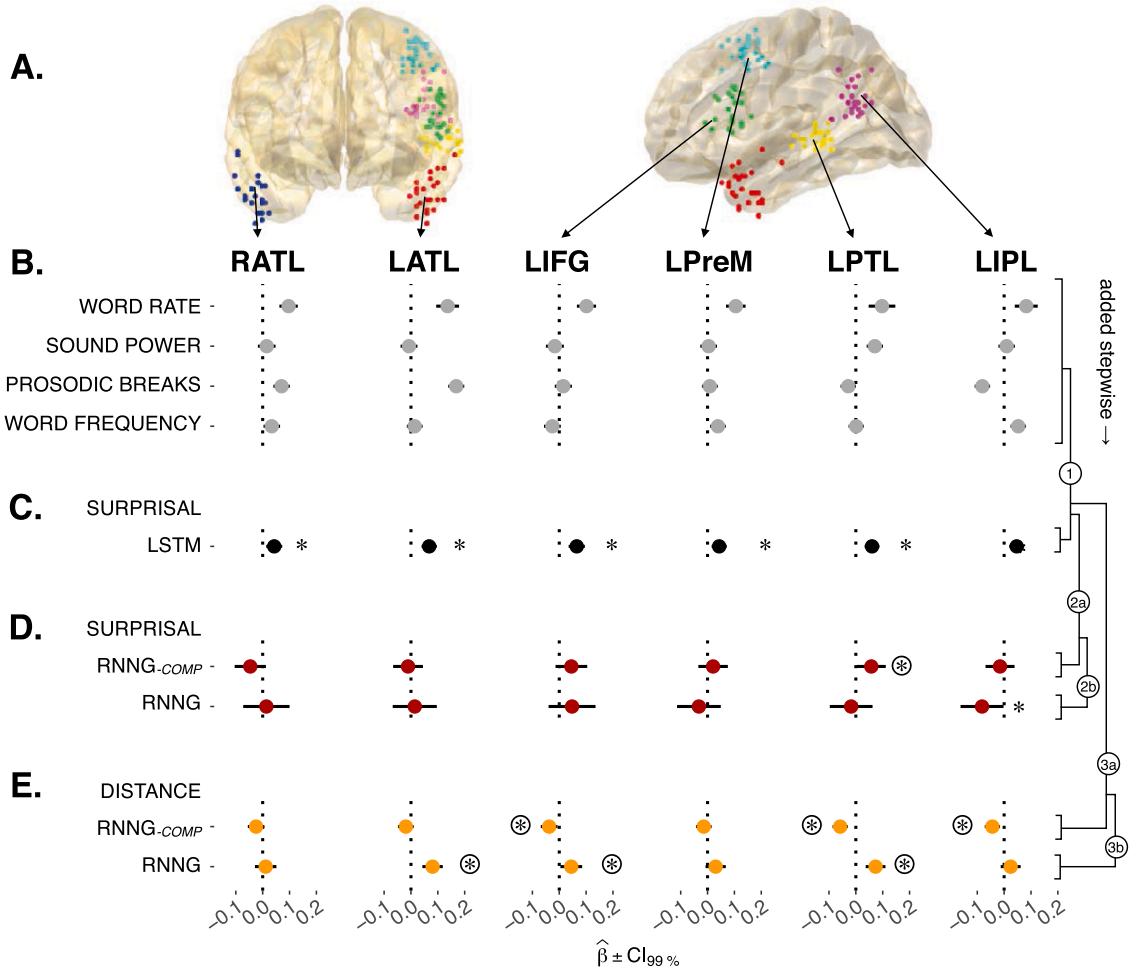


Fig. 2. (A) Individually-defined ROIs for six color-coded regions. (B) Regression coefficients for linguistic control predictors. (C) Regression coefficients for LSTM SURPRISAL when included in a model with all control predictors. (D) Regression coefficients for SURPRISAL from RNNG$_{-COMP}$ and from the full RNNG when added, step-wise, to a model with LSTM SURPRISAL and control predictors. (E) Regression coefficients for DISTANCE from RNNG$_{-COMP}$ and from the full RNNG when added, step-wise, to a model with LSTM SURPRISAL and control predictors. Asterisks (*) indicate a model comparison result such that adding just the indicated term to a model comprised of lower-order terms improves the likelihood of the data with $p < 0.00\overline{83}$, correcting for multiple comparisons across ROIs. Circled asterisks (⊛) indicate such an effect in the presence of a statistically reliable interaction between ROI and the target term. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Stack contents for the example sentence beginning *The rabbit hole …* from the RNNG and RNNG$_{-COMP}$ models along-side an equivalent LSTM representation. Vertical lines delimit distinct stack cells; ▷ indicates boundaries in a word sequence.

RNNG	S	(NP the rabbit hole)					VP	
RNNG$_{-COMP}$	(S	(NP	the	rabbit	hole	$)_{NP}$	(VP	
LSTM	the ▷ rabbit ▷ hole							

2017; Pallier et al., 2011; Snijders et al., 2009; Zaccarella et al., 2017). While prior work has not reached a consensus, the principle loci of such combinatoric processing in that literature include the LATL, LPTL, and the LIFG; we predict that DISTANCE will modulate activity within that network and further expect improved fit to the fMRI data when composition is explicitly encoded.

Earlier work with the same dataset (Brennan et al., 2016) used a set of computational models and complexity metrics that are theoretically related to the LSTM and RNNG models used here, but differ in several key respects. Importantly, those previous models were limited as accounts of comprehension in that they were defined only over parts-of-speech, not individual lexical items, and they did not address the issue of structural ambiguity. We discuss those metrics in more detail in the Supplemental Material, and present analyses incorporating those metrics along-side the RNNG-based models used here.

3. Statistical analysis

We use hierarchical regression and model comparison to test the fit between SURPRISAL and DISTANCE from this family of computational models and the fMRI time-series from six language-related ROIs.

To align the linguistic metrics with the hemodynamic data, we transform the word-by-word values from the computational models, and several control variables that are described below, into estimators for the hemodynamic signal (Brennan et al., 2012, 2016; Just and Varma, 2007). The steps for this transformation are as follows. (i) Each variable of interest (e.g. SURPRISAL, DISTANCE, etc.) is given a time-stamp based on the offset of each word in the audiobook story. (ii) These time-aligned vectors are then convolved with the canonical Hemodynamic Response Function (HRF) using the `spm_volterra()` function from

the SPM8 toolbox and then (iii) orthogonalized against the convolved output for a simple "word rate" indicator variable that is set to one at word-offset and to zero otherwise; this removes low-level correlations between the vectors introduced by their common timing. Finally, (iv) the resulting vectors are re-sampled to 0.5 Hz and truncated to match the 362 samples of fMRI data recorded for each ROI. The results of this procedure are a set of estimators for hemodynamic signals that vary in accordance with states of our computational models as quantified by the complexity metrics.

We constructed separate regression models for the fMRI time-series, *z*-transformed, from each of the six ROIs. Control predictors in the regression models include the convolved vectors for the rate of word occurrence which was also used as a functional language localizer (*z*-transformed), acoustic sound power (*z*-transformed), six movement parameters for each participant estimated during fMRI pre-processing (mean centered), word frequency (HAL corpus values, log-transformed), and a manual annotation of prosodic break strength. (The control predictors were not orthogonalized against the "word rate" indicator variable.) Each of these control factors was entered as a fixed effect into a hierarchical regression model fit with the `lmer()` function from the *lme4* package in R (Bates and Maechler, 2009; R Development Core Team, 2006). A random intercept term was included for each participant and a random slope term was included by participant for the word rate predictor.

Against these baseline models for each ROI, we sequentially add terms of interest derived from the computational models described above. In step (1) we add a term for LSTM SURPRISAL. The analysis then splits into one path for RNNG SURPRISAL and another for RNNG DISTANCE. On one path, we begin with the model with control predictors and the LSTM term and in step (2a) we add the SURPRISAL predictor dervied from RNNG_{-COMP}. Then, further, step (2b) adds both the RNNG_{-COMP} and full RNNG SURPRISAL predictors together. The other path begins also with a model containing all control predictors and the LSTM term, but then adds in step (3a) first just DISTANCE from the RNNG_{-COMP} model, and then in step (3b) both RNNG_{-COMP} and the full RNNG DISTANCE terms together. This step-wise sequence of model-building is schematized on the right-hand side of Fig. 2.

Bivariate correlation coefficients for each term entered into the models are shown in Table 3. Note that LSTM SURPRISAL shows substantial similarity, but not identity, with SURPRISAL from the RNNG models ($r > 0.8$). The metrics derived from the full RNNG, as compared to RNNG_{-COMP}, are also quite similar (both $r > 0.6$); these correlations underscore the importance of step-wise model comparison, described below, to evaluate the unique contribution of each term, and encourage caution when interpreting individual coefficient values.

Constructing models in this step-wise fashion allows us to evaluate the contribution of the RNNG SURPRISAL and DISTANCE terms above-and-beyond lower-order control covariates and LSTM SURPRISAL using model comparison. Model comparsion was done using the `anova()` function in R to compute the likelihood ratio between models that differ in just one term. For nested models, as used here, this ratio follows the χ^2 distribution with degrees of freedom equal to difference in the number of terms between models. Individual terms were considered "statistically significant" if this model comparison reached $p < 0.05/6 = 0.008\overline{3}$, which adjusts the significance threshold for testing across six ROIs using a Bonferroni correction.

4. Results

The six ROIs are shown in different colors in Fig. 2A where each dot represents an individual participant's peak. Regression coefficients from the control model and the step-wise more complex models are plotted, below, in Fig. 2B–F. These coefficients are extracted from the minimal model containing each term. For example, the coefficients for LSTM SURPRISAL in panel C come from the model for each ROI containing just control predictors and LSTM SURPRISAL, but without any terms derived from the RNNG. Statistical significance, indicated by asterisks, comes from step-wise model comparison of the more complex models against simpler models described in the Methods section, above.

Panel B in Fig. 2 shows coefficients for the control predictors. The first row indicates a strong positive effect for the word rate predictor; this result is trivial as the same predictor was originally used to functionally localize these ROIs (Brennan et al., 2016). Results for the other control predictors are discussed below.

Our principal question concerns how linguistic predictions based on different types of syntactic information modulate responses across this set of language-related regions. Panel C in Fig. 2 shows the coefficients in each ROI for LSTM SURPRISAL when this term is included in a model with the control predictors. This coefficient shows a statistically significant positive effect in all six ROIs. Summary statistics for the model comparisons in each ROI are given in Table 4. This result is consistent with prior work showing that lexical predictions based on sentence context modulate activity in a range of fronto-temporal language-related regions as top-down activity "cascades" from higher-level compositional processing through to lower lexical and sub-lexical levels.

We next test in what way the explicit encoding of linguistic hierarchy might modulate activity above-and-beyond the effect of the LSTM. Panel D in Fig. 2 shows the coefficents for SURPRISAL from both the RNNG_{-COMP} and full RNNG models. These are added sequentially, one at a time, onto lower-order models that include LSTM SURPRISAL and control predictors. We find that the RNNG does capture variance in the LPTL and LIPL ROIs as shown in the statistical summaries in Table 5. In the LPTL, RNNG_{-COMP} SURPRISAL captures variance such that higher SURPRISAL leads to greater activity. A different pattern is seen in the LIPL such that SURPRISAL from the full RNNG captures variance above-and-beyond that explained with the LSTM, RNNG_{-COMP}, and control predictors. But, the effect direction is such that there is lower activity for higher SURPRISAL. Such a result should be interpreted with great caution; an important factor to consider is that the LSTM, RNNG_{-COMP}, and full RNNG predictors are highly correlated with each other (see Table 3); this may lead to unstable estimates for the resulting coefficients but does not impact the model comparison statistics. Indeed, when RNNG_{-COMP} and LSTM SURPRISAL are removed from the model, we see a positive coefficient for RNNG SURPRISAL, $\beta = 0.025, SE = 0.010$.

Finally, the DISTANCE metric that we extract from the RNNG allows us to investigate how the number of parse steps explored by the model modulates brain activity across these regions. These effects are shown in panel E of Fig. 2 and model comparisons are summarized in Table 6. RNNG DISTANCE captures variance above-and-beyond LSTM SURPRISAL and other control covariates in the LATL, LIFG, LPTL and the LIPL. Furthermore, these effects were specific to the full RNNG, not RNNG_{-COMP}, in the LATL, LIFG and LPTL.

The results thus far reflect regression analyses that were fit individually to each ROI. We quantify the statistical reliability of differences between regions in the following way. For each of five target terms we construct a regression model which includes that target term and all control and lower-order predictors (just as in the step-wise model comparison) and then we also add a main effect of ROI as well as interactions between ROI and every other term. This model also includes a random slope for the effect of ROI by subject. We then conduct a likelihood ratio test for the change in likelihood when we remove just the interaction term between the target predictor (e.g. RNNG DISTANCE) and ROI.

This interaction analysis indicates that there is no significant interaction between LSTM SURPRISAL and ROI ($\chi^2(5) = 4.96$, $p = 0.421$). There is a statistically significant interaction between RNNG_{-COMP} SURPRISAL and ROI ($\chi^2(5) = 20.37$, $p = 0.001$), between RNNG_{-COMP} DISTANCE and ROI ($\chi^2(5) = 13.15$, $p = 0.022$), and also between RNNG DISTANCE and ROI ($\chi^2(5) = 18.44$, $p = 0.002$). There is no significant interaction with ROI for RNNG SURPRISAL ($\chi^2(5) = 9.09$, $p = 0.105$).

Table 3
Bivariate correlations (Pearson's *r*) between terms entered as predictors into the hierarchical regression models (excluding motion parameters). Off-diagonal cells where $|r| \geq 0.2$ are indicated in **bold face**.

	sound power	word rate	word frequency	prosodic breaks	LSTM surprisal	RNNG distance	RNNG surprisal	RNNG distance	RNNG$_{-COMP}$ surprisal
sound	1.00								
rate	**0.45**	1.00							
frequency	0.09	**0.28**	1.00						
breaks	−0.13	−0.18	−0.17	1.00					
LSTM.s	0.02	−0.03	−0.12	0.15	1.00				
RNNG.d	**−0.22**	0.05	0.10	−0.02	0.08	1.00			
RNNG.s	−0.07	−0.05	−0.18	**0.22**	**0.82**	**0.22**	1.00		
RNNG-c.d	−0.08	0.10	0.04	**−0.20**	0.12	**0.64**	0.17	1.00	
RNNG-c.s	−0.03	−0.06	−0.19	0.13	**0.86**	**0.20**	**0.93**	**0.20**	1.00

Table 4
Model comparison statistics for LSTM SURPRISAL when added step-wise to a model containing all control predictors. Asterisks (*) indicate an improvement in model fit that surpasses a Bonferroni-corrected $\alpha = 0.05/6 = 0.008\overline{3}$.

ROI	LogLik	$\chi^2(1)$	*p*	
LATL	−11530	39.72	<0.001	*
RATL	−11210	14.50	<0.001	*
LIFG	−10649	32.4	<0.001	*
LPTL	−11961	33	<0.001	*
LIPL	−12003	20	<0.001	*
LPreM	−12137	16.5	<0.001	*

Table 5
Model comparison statistics for RNNG and RNNG$_{-COMP}$ SURPRISAL. $\emptyset_{LSTM}$ denotes a model fit with all control predictors as well as LSTM SURPRISAL. Asterisks (*) indicate an improvement in model fit that surpasses a Bonferroni-corrected $\alpha = 0.05/6 = 0.008\overline{3}$.

ROI	Comparison	LogLik	$\chi^2(1)$	*p*
LATL	RNNG$_{-COMP} > \emptyset_{LSTM}$	−11530	0.27	0.60
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−11529	0.21	0.65
RATL	RNNG$_{-COMP} > \emptyset_{LSTM}$	−11208	4.4	0.04
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−11208	0.2	0.67
LIFG	RNNG$_{-COMP} > \emptyset_{LSTM}$	−10646	4.1	0.04
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−10645	2.1	0.15
LPTL	RNNG$_{-COMP} > \emptyset_{LSTM}$	−11957	8.0	0.005*
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−11957	0.3	0.564
LIPL	RNNG$_{-COMP} > \emptyset_{LSTM}$	−12002	0.6	0.457
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−11999	7.3	0.007*
LPreM	RNNG$_{-COMP} > \emptyset_{LSTM}$	−12136	1.0	0.3
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−12136	1.1	0.3

These interaction effects lend support to the claim that the hierarchical effects for RNNG$_{-COMP}$ SURPRISAL and for DISTANCE are specific to certain ROIs.

Returning to the control predictors, the coefficients in Fig, 2B show that sound power has a positive effect in the LPTL, adjacent to the primary auditory cortex. The index of prosodic break strength shows positive effects in both the left and right ATL, and a negative effect in the LIPL. Word frequency shows more moderate effects across regions likely due its reduced impact in a story-book setting with strong contextual influences.

The results discussed above all reflect the step-wise addition of higher-level syntactic predictors over baseline control terms, as illustrated on the right-hand side of Fig. 2. All of the coefficients from each of the models that were fit by this step-wise procedure are summarized in Fig. S1 in the Supplementary Material. To check how sensitive our conclusions might be to the use of step-wise comparison, we also fit a single regression model with all target terms simultaneously. The results

Table 6
Model comparison statistics for RNNG and RNNG$_{-COMP}$ DISTANCE. $\emptyset_{LSTM}$ denotes a model fit with all control predictors as well as LSTM SURPRISAL. Asterisks (*) indicate an improvement in model fit that surpasses a Bonferroni-corrected $\alpha = 0.05/6 = 0.008\overline{3}$.

ROI	Comparison	LogLik	$\chi^2(1)$	*p*
LATL	RNNG$_{-COMP} > \emptyset_{LSTM}$	−11528	3	0.08
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−11513	31	0.001*
RATL	RNNG$_{-COMP} > \emptyset_{LSTM}$	−11208	4.4	0.04
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−11208	0.5	0.46
LIFG	RNNG$_{-COMP} > \emptyset_{LSTM}$	−10643	10.1	0.001*
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−10639	8.5	0.004*
LPTL	RNNG$_{-COMP} > \emptyset_{LSTM}$	−11947	29	< 0.001*
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−11933	28	< 0.001*
LIPL	RNNG$_{-COMP} > \emptyset_{LSTM}$	−11994	16	< 0.001*
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−11993	3	0.08
LPreM	RNNG$_{-COMP} > \emptyset_{LSTM}$	−12136	1.6	0.20
	RNNG > RNNG$_{-COMP} + \emptyset_{LSTM}$	−12134	4.5	0.03

of this fit are given in Fig. 3 (see also Supplemental Table S1). Results for all DISTANCE measures are qualitatively unchanged (yellow points), as is the effect of RNNG$_{-COMP}$ SURPRISAL observed in the LPTL and LIPL (red points). LSTM SURPRISAL shows a reliable positive effect only in the LATL, RATL, and LIPL (black points), but this term does not show reliable effects in the other ROIs when entered along-side RNNG SURPRISAL terms.

A reviewer raised the question of how these results compare with those from an earlier analysis by our group (Brennan et al., 2016) that compared a different set of complexity metrics with the same fMRI data. These include surprisal metrics derived using different language models and also node count metrics based on the number of phrase-structure nodes using a gold-standard syntactic analysis that does not take into account structural ambiguity. Interpreting such a comparison is challenging as the models used in that work differ from the present in several important ways. In addition to lacking an account of structural ambiguity, those models were defined over each word's part-of-speech, while the present models are defined over individual lexical items themselves. The comparison analysis is detailed in the Supplemental Materials (Tables S3, S4, and S5.) Briefly, while RNNG DISTANCE results are unchanged when considered above-and-beyond both surprisal and structure-based node counts, it appears that the prior SURPRISAL predictors capture similar variance to that captured by the RNNG$_{-COMP}$ SURPRISAL predictor.

5. Discussion

This study aims to narrow down the kind of information used to guide predictive processing across different brain regions that are engaged in sentence-level language comprehension. We combined publicly available fMRI data collected while participants listen to a

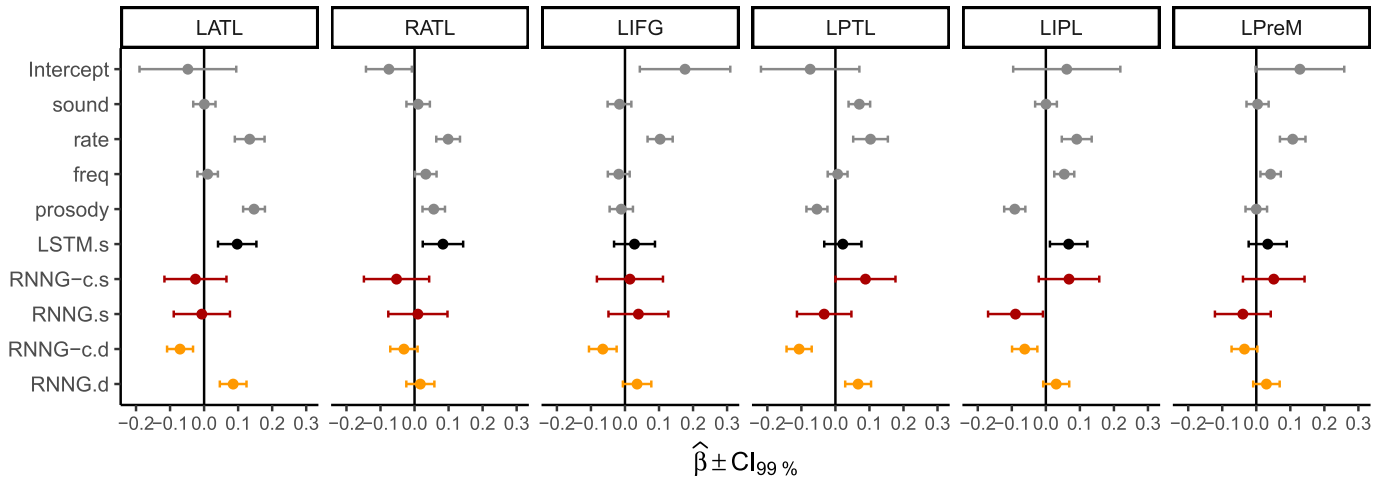


Fig. 3. Regression coefficients with 99% Confidence Interals for all terms (excluding movement parameters) from each of the six ROIs when all terms are fit together in a single model. Abbreviations used: '.s' indicates SURPRISAL measures; '.d' indicates DISTANCE measures; 'RNNG-c' indicates RNNG$_{-COMP}$ models.

naturalistic story with a computational parsing model that we quantify in two ways: the (un)expectedness of a word, or SURPRISAL, and the number of parsing steps, or DISTANCE, explored by the model between words. We examine the fit between these metrics and fMRI activity across a set of fronto-temporal regions to test where the explicit encoding of phrase-structure captures aspects of the brain signal above-and-beyond word-sequence information and other control covariates. We do see evidence for phrase-structure effects: when SURPRISAL is conditioned by explicit hierarchy, as in the RNNG language model, this improves the fit against fMRI data from the LPTL and LIPL. The DISTANCE metric correlates with activity in the LIFG, LATL, LIPL and LPTL and, further all of these fits except for the LIPL are improved when the RNNG not only encodes hierarchy, but composes phrases together into single expressions.

We see effects for SURPRISAL across all six ROIs when that value comes from a LSTM language model (Fig. 2 and Table 4) which does not encode syntactic structure explicitly but may recover such information implicitly (Linzen et al., 2016; Wilcox et al., 2019). This is consistent with a range of prior neuroimaging studies that use a similar methodology as ours, but calculate surprisal from other kinds of language-models (Willems et al., 2016; Lopopolo et al., 2017; Henderson et al., 2016). This result is also in-line with that obtained by Brennan et al. (2016) (using the same datasets as analyzed here) who see effects when surprisal is defined by a probabalistic context-free grammar in all six of these same ROIs. The model used in that work, unlike the present model, calculated probabilities for each word's part-of-speech, not the lexical item itself. Henderson et al. (2016) also report effects when surprisal is calculated for parts-of-speech using a phrase structure grammar, and find their largest effect in the LIFG, along-side a less robust effect in the LATL, while other work in which surprisal is defined on the basis of only sequential information (a Markov model) find primarily effects in the temporal lobes bilaterally (Willems et al., 2016; Lopopolo et al., 2017).

When SURPRISAL is calculated from a RNNG language model, which does encode hierarchy explicitly, such a model improves fits to fMRI signals recorded in the LPTL and LIPL (Fig. 2 and Table 5).[4] This pattern holds whether or not the RNNG DISTANCE terms are included in the model fits (Fig. 3). Brennan et al. (2016) also found hierarchical surprisal improved fits in both the LPTL and LIPL above-and-beyond a sequential Markov-model baseline. As already noted above, they characterized hierarchical surprisal using a one-path gold-standard context-free grammar parse-tree and only computed surprisal for each word's part-of-speech, not for the lexical item itself. They also report improved fits of the same kind in other regions including LATL, RATL, and LIFG, which we do not find in the present results. One reason for that difference may be due to the more sophisticated baseline LSTM model used in the present study; its gated memory architecture allows it more flexibility in capturing contextual dependencies that are both proximal and distal to the target word. Unfortunately, the present analysis does not offer a minimal comparison to that earlier work: in addition to different architectures, the earlier models had a different domain (parts-of-speech) than the present models (lexical items).

The results for SURPRISAL are not statistically reliable when tested along-side SURPRISAL effects derived from models used by Brennan et al. (2016) (Supplemental Table S4; the LIPL effect reaches $p = 0.01$, which does not meet our bonferroni-corrected threshold of $\alpha = 0.008\overline{3}$). This indicates that the two kinds of models are capturing shared variance in the fMRI data. But, understanding the nature of any overlap is challenging due to the already-mentioned differences in their domain and other architectural aspects. The fact that language-users themselves predict specific lexical items, and do so in light of the regular ambiguity in every-day language, recommends something with the same features as the present RNNG models *a priori*. Yet, the empirical fits for part-of-speech-based predictions from those earlier, and simpler, models suggests to us that it will be valuable in future work to probe how to integrate part-of-speech predictions along-side lexical-specific predictions.

Along-side effects for SURPRISAL, we use the DISTANCE complexity metric to probe for neural signals reflecting compositional processing. DISTANCE counts the steps of structural analysis pursued by the parser as it moves from word-to-word. In this way, it aligns with a large number of prior efforts that probe for composition effects by comparing stimuli with more or less syntactic structure, including sentences and word-lists (e.g. Mazoyer et al., 1993; Stowe et al., 1998; Vandenberghe et al., 2002; Humphries et al., 2006; Rogalsky and Hickok, 2009; Brennan and Pylkkänen, 2012; Pallier et al., 2011), simple phrases to non-phrasal stimuli (e.g. Pylkkänen et al., 2014; Bemis and Pylkkänen, 2011; Blanco-Elorrieta and Pylkkänen, 2016; Westerlund et al., 2015), or via computational parsing models (Brennan et al., 2012; Brennan and Pylkkänen, 2017; Nelson et al., 2017b). The effects we see of DISTANCE in

[4] The effect for RNNG SURPRISAL in the LIPL is in the negative direction. This directionality appears to be influenced by the presence of highly correlated predictors like RNNG$_{-COMP}$ surprisal. We do not here speculate on the interpretation of this direction because the combination of data and models appear insufficient to accurately estimate its direction when taking lower-order predictors into account.

the LATL, LPTL, LPTL and LIFG are broadly in accordance with results from those prior studies (Fig. 2 and Table 6).[5] Furthermore, the model comparison between the full RNNG DISTANCE values, and those values when the RNNG does not explicitly compose phrasal units together (RNNG$_{-COMP}$) in the LATL, LIFG, and LPTL lends support to the cognitive importance of compositional mechanisms during online parsing (Lewis and Phillips, 2015; Sprouse and Hornstein, 2016; Hale et al., 2018). Finally the results for DISTANCE are not sensitive to whether SURPRISAL terms are or are not included in the models (Fig. 3), nor are they sensitive to the presence of node-counting measures that were used to estimate structural analysis effort in earlier work (Supplemental Table S5). What sets DISTANCE apart from those earlier measures is that it explicitly reflects the ambiguity present in every-day parsing and quantifies structural analysis in the presence of this ambiguity.

A possible point of contention concerns the relative breadth of our DISTANCE result as compared to more narrow findings reported in some prior studies, especially those focused on minimal comparisons of simple phrases to non-phrases (e.g. Bemis and Pylkkänen, 2011, et seq.). Those studies do not find effects in inferior frontal regions. We suggest that careful consideration of the DISTANCE metric can help to resolve the apparent mismatch. The observed correlations may reflect a neural mechanism that varies in proportion to the number of parser operations – in accordance with a kind of semantic composition or syntactic structure-building. But, the parser tends to take more actions when it has entered states that are unexpected or unusual; in other words, when new information requires a kind of syntactic reanalysis. Indeed, several prominent theories of the role for the LIFG in language comprehension assign a role related to addressing ambiguous or conflicted input (e.g. Bornkessel-Schlesewsky and Schlesewsky, 2013; Novick et al., 2010). The suggestion that DISTANCE may relate, at least in part, to late-stage reanalysis operations is supported by results from Hale et al. (2018) who find this measure to correlate with a late, P600-like, EEG component. Future work using more granular complexity metrics to tease apart these different aspects of the DISTANCE metric may be useful in offering a more nuanced window into this particular pattern of results.

6. Conclusion

The expectedness of a word modulates brain activity in a wide range of regions associated with many levels of language comprehension. Using fMRI data from story-listening and a parser based on Recurrent Neural Network Grammars, we see evidence that explicit representation of hierarchy plays a role in a subset of these regions, like the left posterior temporal lobe. Parsing steps that not only encode structure, but compose phrases together into single terms, modulates activity in the left temporal lobe and inferior frontal regions. These findings generalize across the range of linguistic input found in a natural literary text. These quantitative matches between states of a computational model and neural data narrow down the kinds of information that are being processed across perisylvian language-related bain regions. These results point towards future work that probes, in a more granular way, the specific hierarchical structures that are in play during incremental processing.

Acknowledgements

This work was funded in part by grants from the National Science Foundation grants #1607251 (JRB) and #1607441 (JTH).

[5] The direction of the RNNG$_{-COMP}$ DISTANCE effect is negative. But, as already noted for the RNNG$_{-COMP}$ surprisal results, this may be influenced by the high correlation between the relevant measures (see Table 3). As before, because of this we do not speculate about the interpretation of the directionality of this effect.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2020.107479.

References

Bates, D., Maechler, M., 2009. lme4: linear mixed-effects models using S4 classes. R package version 0.999375-31.
Bemis, D.K., Pylkkänen, L., 2011. Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. J. Neurosci. 31 (8), 2801–2814.
Blanco-Elorrieta, E., Pylkkänen, L., 2016. Composition of complex numbers: delineating the computational role of the left anterior temporal lobe. Neuroimage 124 (Pt A), 194–203.
Bornkessel-Schlesewsky, I., Schlesewsky, M., 2013. Reconciling time, space and function: a new dorsal-ventral stream model of sentence comprehension. Brain Lang. 125 (1), 60–76.
Boston, M.F., Hale, J., Kliegl, R., Patil, U., Vasishth, S., 2008. Parsing costs as predictors of reading difficulty: an evaluation using the Potsdam Sentence Corpus. Journal of Eye Movement Research 2 (1), 1–12.
Brennan, J.R., Hale, J.T., 2019. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. PloS One 14 (1), e0207741.
Brennan, J.R., Lajiness-O'Neill, R., Bowyer, S.M., Kovelman, I., Hale, J.T., 2018. Predictive sentence comprehension during story-listening in autism spectrum disorder. Language, Cognition, and Neuroscience 34 (4), 428–439.
Brennan, J.R., Nir, Y., Hasson, U., Malach, R., Heeger, D.J., Pylkkänen, L., 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. Brain Lang. 120, 163–173.
Brennan, J.R., Pylkkänen, L., 2012. The time-course and spatial distribution of brain activity associated with sentence processing. Neuroimage 60, 1139–1148.
Brennan, J.R., Pylkkänen, L., 2017. MEG evidence for incremental sentence composition in the anterior temporal lobe. Cognit. Sci. 41 (S6), 1515–1531.
Brennan, J.R., Stabler, E.P., Van Wagenen, S.E., Luh, W.-M., Hale, J.T., 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. Brain Lang. 157–158, 81–94.
Caplan, D., 1992. Language: Structure, Processing, and Disorders. MIT Press, Cambridge, MA.
Demberg, V., Keller, F., 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition 101 (2), 193–210.
Dikker, S., Pylkkänen, L., 2012. Predicting language: MEG evidence for lexical preactivation. Brain Lang. 127 (1), 55–64.
Dyer, C., Ballesteros, M., Ling, W., Matthews, A., Smith, N.A., 2015. Transition-based dependency parsing with stack long short-term memory. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing Volume 1, 334–343, 08075.
Dyer, C., Kuncoro, A., Ballesteros, M., Smith, N.A., 2016. Recurrent neural network grammars. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, p. 199. –209.
Frank, S.L., Bod, R., 2011. Insensitivity of the human sentence-processing system to hierarchical structure. Psychol. Sci. 22 (6), 829–834.
Frank, S.L., Otten, L.J., Galli, G., Vigliocco, G., 2015. The ERP response to the amount of information conveyed by words in sentences. Brain Lang. 140, 1–11.
Gibson, E.E.F., 1991. A Computational Theory Of Human Linguistic Processing: Memory Limitations And Processing Breakdown. PhD Thesis. Carnegie Mellon University.
Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., Baroni, M., 2018. Colorless green recurrent networks dream hierarchically. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Association for Computational Linguistics, pp. 1195–1205.
Hale, J., 2001. A probabilistic Earley parser as a psycholinguistic model. In: North American Chapter of the Association for Computational Linguistics, vols. 1–8. Association for Computational Linguistics Morristown, NJ, USA.
Hale, J., Dyer, C., Kuncoro, A., Brennan, J., 2018. Finding syntax in human encephalography with beam search. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1. Association for Computational Linguistics, pp. 2727–2736.
Hale, J.T., 2014. Automaton Theories of Human Sentence Comprehension. CSLI Publications.
Hale, J.T., 2016. Information-theoretical complexity metrics. Language and Linguistics Compass 10 (9), 397–412.
Henderson, J.M., Choi, W., Lowder, M.W., Ferreira, F., 2016. Language structure in the brain: a fixation-related fmri study of syntactic surprisal in reading. Neuroimage 132, 293–300.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Humphries, C., Binder, J.R., Medler, D.A., Liebenthal, E., 2006. Syntactic and semantic modulation of neural activity during auditory sentence comprehension. J. Cognit. Neurosci. 18 (4), 665–679.
Jelinek, F., 1998. Statistical Methods for Speech Recognition. MIT Press, Cambridge MA.
Jurafsky, D., 1996. A probabilistic model of lexical and syntactic access and disambiguation. Cognit. Sci. 20 (2), 137–194.

Just, M.A., Varma, S., 2007. The organization of thinking: what functional brain imaging reveals about the neuroarchitecture of complex cognition. Cognit. Affect Behav. Neurosci. 7 (3), 153–191.
Kaplan, R.M., 1972. Augmented transition networks as psychological models of sentence comprehension. Artif. Intell. 3, 77–100.
Klein, D., Manning, C.D., 2003. Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-, Volume 1. Association for Computational Linguistics, pp. 423–430.
Kuncoro, A., Ballesteros, M., Kong, L., Dyer, C., Neubig, G., Smith, N.A., 2016. What do recurrent neural network grammars learn about syntax? Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics Volume 1, 1249–1258, 05774, abs/1611.
Kutas, M., Federmeier, K.D., 2000. Electrophysiology reveals semantic memory use in language comprehension. Trends Cognit. Sci. 4, 463–469.
Lewis, S., Phillips, C., 2015. Aligning grammatical theories and language processing models. J. Psycholinguist. Res. 44 (1), 27–46.
Linzen, T., Dupoux, E., Goldberg, Y., 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Transactions of the Association for Computational Linguistics 4, 521–535.
Lopopolo, A., Frank, S.L., van den Bosch, A., Willems, R.M., 2017. Using stochastic language models (slm) to map lexical, syntactic, and phonological information processing in the brain. PloS One 12 (5), e0177794.
Lowder, M., Choi, W., Ferreira, F., Henderson, J., 2018. Lexical predictability during natural reading: effects of surprisal and entropy reduction. Cognit. Sci. 42, 1166–1183.
Marcus, M.P., Santorini, B., Marcinkiewicz, M.A., 1993. Building a large annotated corpus of English: the Penn Treebank. Comput. Ling. 19, 313–330.
Marslen-Wilson, W.D., 1975. Sentence perception as an interactive parallel process. Science 189 (4198), 226–228.
Mazoyer, B.M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L., Mehler, J., 1993. The cortical representation of speech. J. Cognit. Neurosci. 5 (4), 467–479.
Mikolov, T., Karafiát, M., Burget, L., Černocká, J.H., Khudanpur, S., 2010. Recurrent neural network based language model. Proceedings of Interspeech 2010 1045–1048.
Molinaro, N., Barraza, P., Carreiras, M., 2013. Long-range neural synchronization supports fast and efficient reading: eeg correlates of processing expected words in sentences. Neuroimage 72, 120–132, 0.
Nelson, M.J., Dehaene, S., Pallier, C., Hale, J.T., 2017. Entropy reduction correlates with temporal lobe activity. In: Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics. Association for Computational Linguistics, pp. 1–10.
Nelson, M.J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S.S., Naccache, L., Hale, J.T., Pallier, C., Dehaene, S., 2017b. Neurophysiological dynamics of phrase-structure building during sentence processing. Proc. Natl. Acad. Sci. U. S. A. 114 (18), E3669–E3678.
Nieuwland, M., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D.J., Rousselet, G., Ferguson, H.J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E.M., Donaldson, D.L., Kohút, Z., Rueschemeyer, S.-A., Huettig, F., 2018. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. eLife 7 (e33468).
Novick, J.M., Trueswell, J.C., Thompson-Schill, S.L., 2010. Broca's area and language processing: evidence for the cognitive control connection. Language and Linguistics Compass 4 (10), 906–924.
Pallier, C., Devauchelle, A.-D., Dehaene, S., 2011. Cortical representation of the constituent structure of sentences. Proc. Natl. Acad. Sci. Unit. States Am. 108 (6), 2522–2527.
Pylkkänen, L., 2016. Composition of complex meaning: interdisciplinary perspectives on the left anterior temporal lobe. In: Hickok, G., Small, S. (Eds.), Neurobiology of Language. Academic Press, London.
Pylkkänen, L., Bemis, D.K., Blanco Elorrieta, E., 2014. Building phrases in language production: an meg study of simple composition. Cognition 133 (2), 371–384.
R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Roark, B., 2004. Robust garden path parsing. Nat. Lang. Eng. 10 (1), 1–24.
Roark, B., Bachrach, A., Cardenas, C., Pallier, C., 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 324–333.
Rogalsky, C., Hickok, G., 2009. Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. Cerebr. Cortex 19 (4), 786–796.
Snijders, T.M., Vosse, T., Kempen, G., Van Berkum, J.J.A., Petersson, K.M., Hagoort, P., 2009. Retrieval and unification of syntactic structure in sentence comprehension: an fmri study using word-category ambiguity. Cerebr. Cortex 19 (7), 1493–1503.
Sprouse, J., Hornstein, N., 2016. Syntax and the cognitive neuroscience of syntactic structure building. In: Hickok, G., Small, S.L. (Eds.), Neurobiology of Language. Academic Press, pp. 165–174 (chapter 14).
Stern, M., Fried, D., Klein, D., 2017. Effective inference for generative neural parsing. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1695–1700.
Stowe, L.A., Broere, C.A., Paans, A.M., Wijers, A.A., Mulder, G., Vaalburg, W., Zwarts, F., 1998. Localizing components of a complex task: sentence processing and working memory. Neuroreport 9 (13), 2995–2999.
Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., Sedivy, J., 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268 (5217), 1632–1634.
Vandenberghe, R., Nobre, A.C., Price, C.J., 2002. The response of left temporal cortex to sentences. J. Cognit. Neurosci. 14 (4), 550–560.
Westerlund, M., Kastner, I., Al Kaabi, M., Pylkkänen, L., 2015. The LATL as locus of composition: MEG evidence from English and Arabic. Brain Lang. 141, 124–134.
Wilcox, E., Levy, R., Futrell, R., 2019. Hierarchical representation in neural language models: suppression and recovery of expectations. Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP 181–190, 2019.
Willems, R.M., Frank, S.L., Nijhof, A.D., Hagoort, P., van den Bosch, A., 2016. Prediction during natural language comprehension. Cerebr. Cortex 6 (1).
Zaccarella, E., Meyer, L., Makuuchi, M., Friederici, A.D., 2017. Building by syntax: the neural basis of minimal linguistic structures. Cerebr. Cortex 27 (1), 411–421.