Using surprisal and fMRI to map the neural bases of broad and local contextual prediction during natural language comprehension

Shohini Bhattasali and **Philip Resnik**
Linguistics/UMIACS
University of Maryland
College Park, MD
{shohini, resnik}@umd.edu

Abstract

Context guides comprehenders' expectations during language processing, and information-theoretic surprisal is commonly used as an index of cognitive processing effort. However, prior work using surprisal has considered only within-sentence context, using n-grams, neural language models, or syntactic structure as conditioning context. In this paper, we extend the surprisal approach to use broader topical context, investigating the influence of local and topical context on processing via an analysis of fMRI time courses collected during naturalistic listening. Lexical surprisal calculated from ngram and LSTM language models is used to capture effects of local context; to capture the effects of broader context a new metric based on topic models, topical surprisal, is introduced. We identify distinct patterns of neural activation for lexical surprisal and topical surprisal. These differing neuro-anatomical correlates suggest that local and broad contextual cues during sentence processing recruit different brain regions and that those regions of the language network functionally contribute to processing different dimensions of contextual information during comprehension. More generally, our approach adds to a growing literature using methods from computational linguistics to operationalize and test hypotheses about neuro-cognitive mechanisms in sentence processing.

1 Introduction

Narratives unfold over time and comprehenders incrementally process words and sentences. In order to understand the current word and sentence, we have to integrate current input with the information from the previous context.

In characterizing this process, the notion of *surprisal* from information theory has been prevalent in psycholinguistic modeling, following the work of Hale (2001) and Levy (2008). Surprisal operationalizes how unexpected a word is as its pointwise information content given prior context, $-\log \mathrm{P}(w_i|w_1...w_{i-1})$. Generally the theory of surprisal, as applied in the study of human language comprehension, proposes that probabilistic predictions made by comprehenders yield variability in word-by-word processing difficulty: when surprisal is high, the current word is unexpected and cognitive processing effort increases accordingly. This linking has been validated in prior work connecting surprisal with measurable reflexes of cognitive effort, using probabilities conditioned on lexical (sequential) and syntactic contexts (e.g. Brennan et al., 2016; Lopopolo et al., 2017; Brennan and Hale, 2019; Shain et al., 2020).

In this paper, we extend the surprisal paradigm beyond prior work looking only at local context, to investigate the influence of broader contextual information during incremental sentence processing. As an illustration, consider examples (a) and (b), which illustrate that, while a word might be extremely difficult to predict given the immediate local context, it might be less unexpected/more predictable given the broader topic under discussion. The actual word to be predicted here is *China*, which is not at all predictable given the immediate context, but more likely in a longer discourse about traveling (*airplanes*, *places*, *world*, *geography*, etc).

(a) I could recognize at first glance ______

(b) *So I had to choose another profession, and I learned to fly airplanes. I flew a little in many places around the world. And geography it's true has served me well.* I could recognize at first glance ______

Following prior studies, we use measures of lexical surprisal to capture the influence of local context, and we introduce a new measure, *topical*

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3786–3798
August 1–6, 2021. ©2021 Association for Computational Linguistics

surprisal, to operationalize the predictive role of broader context. We use both kinds of predictors to investigate how processing based on different contextual cues differs in their mapping in the brain. Specifically, our research questions are:

- How does the previous topical context affect our expectations about the next word?
- How do local vs. broad contextual prediction influence our incremental language processing? Do they have distinct neural correlates?

We apply computational modeling to investigate these questions in a way that would not be feasible in a more traditional, trial-based experimental paradigm by taking advantage of data collected using fMRI brain imaging during continuous, naturalistic listening (Hamilton and Huth, 2020). This data collection method has emerged as a new testing ground for linking processing hypotheses with neurobiological architectures in the brain (Maguire, 2012; Willems, 2015; Kandylaki and Bornkessel-Schlesewsky, 2019). Using lexical surprisal and our new measure of topical surprisal as computational predictors of cognitive activity, we demonstrate that processing of local and broad context recruits different brain regions, suggesting that those regions of the language network functionally contribute to processing different dimensions of contextual prediction during human language comprehension.

2 Background and Related Work

2.1 Surprisal as a cognitive measure

Prior neurolinguistic work has used surprisal as an index of cognitive processing effort. Behavioral measurements like self-paced reading are one way to infer how much effort is involved while processing some piece of linguistic input (e.g., Futrell et al., 2021); other methods more directly measure activity in the brain, including functional magnetic resonance imaging (fMRI), which we will focus on in this paper, as well as magnetoencephalography (MEG) (e.g., Brodbeck et al., 2018) and electroencephalography (EEG) (e.g., Ettinger et al., 2016; Brennan and Hale, 2019; Michaelov and Bergen, 2020).

In such work, the logic is generally as follows. First, as noted in §1 we assume that when a word is less expected given the context, processing it during comprehension will require more work in the brain. Then, we computationally estimate a model of surprisal using a corpus:

$$\text{surprisal}_M(w_i) = -\log \text{P}_M(w_i|w_1...w_{i-1}) \quad (1)$$

Two common instantiations for M include ngram models and models conditioned on prior syntactic context (Hale, 2001).

By the first assumption, the value of Eq (1) is taken to be a predictor of processing effort at word w_i. Therefore, the key final step is to analyze the relationship between that estimated effort, as predicted by the model, and observed activity measured in the brain. In the case of fMRI, neural activity is measured by detecting changes associated with blood flow (see §4). When there are significant correlations between the predicted effort, surprisal$_M$, and activity in some region of the brain, this constitutes evidence for that region being involved in processing of the information that M has used in its predictions. For example, if the brain activity in a region is correlated with an estimate of surprisal that uses syntactic predictions, that provides evidence for that region being a locus for human syntactic processing.

In prior work following this logic, using the surprisal paradigm with fMRI to localize processing associated with lexical and syntactic context, the findings implicate a range of core regions of the language network. Across different languages, lexical surprisal recruits a mostly left-lateralized network, predominantly consisting of Inferior Frontal Gyrus, Interior Temporal Sulcus, Middle Frontal Gyrus, Posterior Temporal regions, extending to some bilateral regions, namely Anterior Temporal Lobe and Superior Temporal Gyrus (Brennan, 2016; Willems et al., 2016; Lopopolo et al., 2017; Shain et al., 2020). Syntactic surprisal has also mapped onto a left-lateralized network consisting of the Inferior Parietal Lobule, Inferior Frontal Gyrus, Middle Temporal Gyrus, along with some evidence for bilateral processing in the Anterior Temporal Lobe (Brennan et al., 2016; Henderson et al., 2016; Lopopolo et al., 2017; Shain et al., 2020).

2.2 Neural language models in cognitive neuroscience

There has been a growing trend of using neural language models in cognitive neuroscience research, often using neural data collected from individuals during naturalistic listening.[1] As one salient example, Wehbe et al. (2014) investigated how well vector representations predicted brain activity for subjects reading fiction, in their case material from *Harry Potter and the Sorcerer's Stone*, based on within-sentence context. Also working within the sentence using naturalistic listening, Toneva et al. (2020) derived composed representations of "supra-word meaning" using contextualized word representations (ELMo, Peters et al., 2018) to capture the compositional meaning of multi-word expressions and event/argument structure. Jain and Huth (2018) make predictions of neural activity using LSTM representations from up to the previous 20 words of context (which would be on the order of 8-10 seconds of speech on average).

Work of this kind has a number of dimensions of variation. One is the nature of the neural measurement, e.g. fMRI versus MEG, which relates crucially to the cognitive questions being asked, since some questions involve temporal locality, a strength of MEG, and others involve spatial locality, a strength of fMRI. Another dimension is the nature of the training data for the computational modeling, e.g. material from the experimental dataset (*Harry Potter*, as in Wehbe et al. (2014)) versus a broader coverage corpus such as a large collection of Reddit comments as used by Jain and Huth (2018)). Finally, there is the nature of the model itself; for example, how much context it takes into account and whether it involves, for example, non-contextual word embeddings, sequentially derived embeddings, or something else.

In this work, we use broad coverage corpora such as COCA (Davies, 2008) or Wikipedia to train our models. In addition to using ngram and LSTM models to capture within-sentence context, we introduce topical surprisal (§3.5), based on topic modeling, as a way to look at functional localization of correlates of broader, non-sequential contextual processing using fMRI.

3 fMRI Study

3.1 Method

We follow Brennan et al. (2012) in using a spoken narrative as a stimulus. Participants hear a story over headphones while they are in the MRI scanner. As we describe in greater detail in §4, the sequence of neuroimages collected during their session becomes the dependent variable in a regression against word-by-word predictors that have been derived from the text of the story.

3.2 Stimuli

The English audio stimulus was Antoine de Saint-Exupéry's *The Little Prince*, translated by David Wilkinson and read by Karen Savage. It constitutes a fairly lengthy exposure to naturalistic language, comprising 19,171 tokens; 15,388 words and 1,388 sentences, and lasting over an hour and a half. *The Little Prince* has been used in a number of previous fMRI studies of language processing, e.g. Li et al. (2018); Bhattasali et al. (2019); Zhang (2020); Stanojević et al. (2021)

3.3 Participants

56 participants were scanned and 5 of them were excluded since they had incomplete scanning sessions. Participants included were fifty-one volunteers (32 women and 19 men, 18-37 years old) with no history of psychiatric, neurological, or other medical illness or history of drug or alcohol abuse that might compromise cognitive functions. All strictly qualified as right-handed on the Edinburgh handedness inventory (Oldfield, 1971). All self-identified as native English speakers and gave their written informed consent prior to participation, in accordance with Cornell University's IRB guidelines. Participants were compensated for their time, consistent with typical practice for studies of this kind. They were paid $65 at the end of the session.

3.4 Presentation

After giving their informed consent, participants were familiarized with the MRI facility and assumed a supine position on the scanner gurney. The presentation script was written in PsychoPy (Peirce, 2007). Auditory stimuli were delivered through MRI-safe, high-fidelity headphones (Confon HP-VS01, MR Confon, Magdeburg, Germany) inside the head coil. The headphones were secured against the plastic frame of the coil using foam blocks. Using a spoken recitation of the US Constitution, an experimenter increased the volume until participants reported that they could hear clearly. Participants then listened passively to the audio storybook for 1 hour 38 minutes. The story had nine chapters and at the end of each chapter the

[1] Although not directly relevant to the scientific strategy we discuss here, we note that there is also a body of work that goes in the other direction, using methods from psycholinguistics and neuroscience to improve our understanding and use of neural language models, e.g. Toneva and Wehbe (2019); Ettinger (2020); Misra et al. (2020).

participants were presented with a multiple-choice questionnaire with four questions (36 questions in total), concerning events and situations described in the story. These questions served to confirm participants' comprehension. They were viewed via a mirror attached to the head coil and answered through a button box. The entire session lasted around 2.5 hours.[2]

3.5 Deriving the predictors

Recall that surprisal measures how unexpected each word w_i is given the preceding context (Eq 1), and we use measures of surprisal as the linking hypotheses in our study between the contextual predictions of our model and neural activity. The three different surprisal predictors we use are described below, along with how they were calculated. Fig. 2 shows a visual comparison of the word-by-word predictors on a single sentence from the text.[3]

Ngram surprisal. The ngram surprisal values are based on a 5gram language model and were calculated using the `kenlm` library (Heafield et al.) with Modified Kneser-Ney Smoothing. The 5gram language model was trained on the Corpus of Contemporary American English (COCA, Davies, 2008), which is a large, genre-balanced corpus of American English and consists of over one billion words of text sampled across spoken, fiction, popular magazines, newspapers, and academic texts.

LSTM surprisal. These surprisal values are based on a long short-term memory (LSTM) language model (Hochreiter and Schmidhuber, 1997) trained on 90 million words of English Wikipedia by Gulordava et al. (2018). It had two LSTM layers with 650 hidden units each, 650 dimensional word embeddings, a learning rate of 20, a dropout rate of 0.2 and a batch size 128, and was trained for 40 epochs (with early stopping). Like the majority of previous work computing LSTM-surprisal, our input is a single sentence and we make predictions only based on context within the sentence (Brennan and Hale (2019); van Schijndel and Linzen (2018), though cf. Jain and Huth (2018)). The surprisal values were calculated using the Neural Complexity toolkit with the baseline non-adaptive model (van Schijndel and Linzen, 2018).

Topical surprisal. We introduce a new predictor based on topic models, adapting surprisal to operationalize the influence of context beyond the sentence level. Topical surprisal for a word is defined as the weighted average of the word's probability given topic, where weights are the (posterior) probability for the topic in that context.

$$\text{surprisal}(w_i \text{ in context c}) = -\log \sum_{t \in \text{Topics}} P(w_i|t)P(t|c) \tag{2}$$

Fig. 1 illustrates how topical surprisal is computed using a sample excerpt from the text.

Topics are defined and probabilities estimated using an LDA topic model (Blei et al., 2003). Using the wrapper for Mallet LDA (McCallum) in the Gensim toolkit (Řehůřek and Sojka), we estimated a 100-topic model with the default hyperparameters using 219,380 documents from COCA, yielding P$(w|t)$ for all word-topic pairs and making it possible to compute the posterior topic probabilities P$(t|c)$ for any new document c. We compute topical surprisal for all the non-function words in the audio sample using the content in the 30-second window prior to the word to define the LDA ”document” c.

4 Data Analysis

fMRI data is acquired with physical, biological constraints and we followed a standard preprocessing pipeline for fMRI imaging data that allowed us to make adjustments to improve the signal to noise ratio.[4]

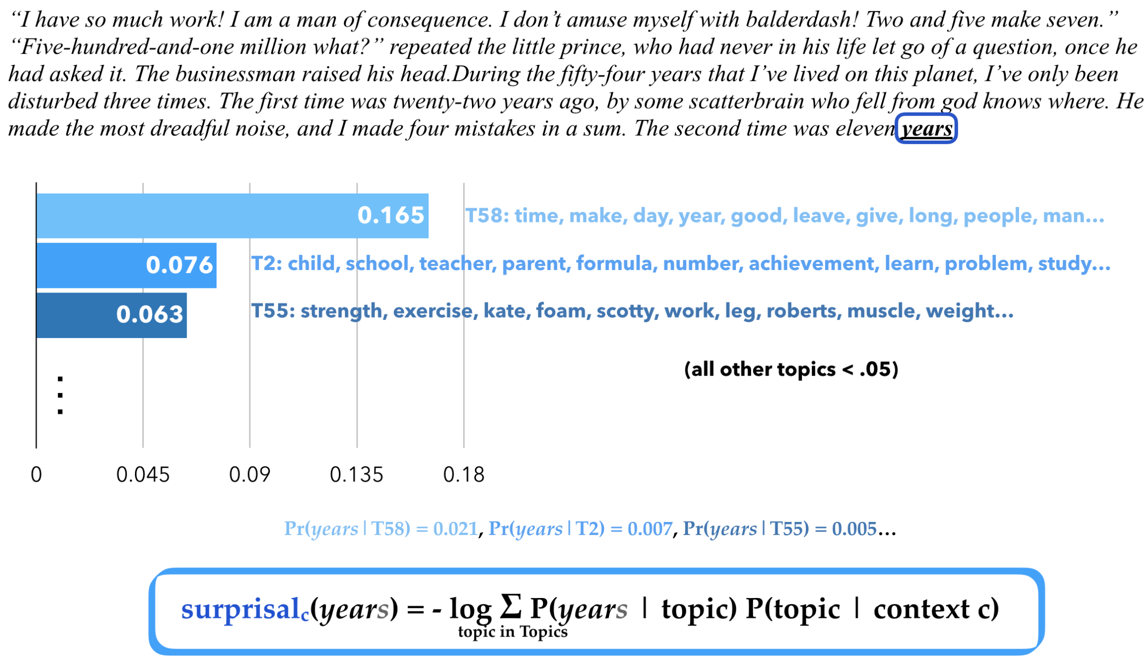
The research questions presented above in §1 motivate two statistical analyses looking at correlations between model-based predictions and observed brain activity. In Analysis 1, we use ngram surprisal (from a 5gram language model) to instantiate local context and compare it against topical surprisal, which captures the influence of broader, topical context. As a follow-up, in Analysis 2, we use surprisal from a state-of-the-art LSTM language model to instantiate local context while still using topical surprisal for broader context, in order to illuminate potential differences between the neural correlates of ngram and LSTM models.

Measurements of neural activity using fMRI are based on an increase in blood flow to regions of the brain, which reflects increased cerebral activity. Because blood flow is slow relative to neural activity, this introduces a temporal lag and presents a

[2] Further details about the fMRI data collection can be found in the Appendix.

[3] A correlation matrix for the predictors is included in the Appendix.

[4] See the Appendix for further preprocessing details.



"I have so much work! I am a man of consequence. I don't amuse myself with balderdash! Two and five make seven." "Five-hundred-and-one million what?" repeated the little prince, who had never in his life let go of a question, once he had asked it. The businessman raised his head.During the fifty-four years that I've lived on this planet, I've only been disturbed three times. The first time was twenty-two years ago, by some scatterbrain who fell from god knows where. He made the most dreadful noise, and I made four mistakes in a sum. The second time was eleven ***years***

0.165
T58: time, make, day, year, good, leave, give, long, people, man...
0.076
T2: child, school, teacher, parent, formula, number, achievement, learn, problem, study...
0.063
T55: strength, exercise, kate, foam, scotty, work, leg, roberts, muscle, weight...
(all other topics < .05)
0
0.045
0.09
0.135
0.18
Pr(years | T58) = 0.021, Pr(years | T2) = 0.007, Pr(years | T55) = 0.005...

$$\text{surprisal}_c(\textit{years}) = -\log \sum_{\text{topic in Topics}} P(\textit{years} \mid \text{topic})\, P(\text{topic} \mid \text{context } c)$$

Figure 1: Real example illustrating the computation of topical surprisal. We calculate the probability of the word *years* conditioned on its topical context by using the previous 30 seconds of of the story to define a "document" c and computing its posterior topic distribution based on an LDA model for a large, diverse collection of English text. In the figure each topic is represented by its highest probability words.

challenge for modeling the time course of processing. To address this issue, predictors are convolved using a canonical hemodynamic response function (HRF) to model the observed time-course of the brain's hemodynamic response (BOLD - Blood Oxygenation Level Dependent) in each voxel.[5] Brennan (2016) provides a detailed description of how word-by-word predictors are convolved to estimate the fMRI BOLD signal in studies like the present one.

In order to look at correlations between predictors and brain activity, our analyses employ the General Linear Model (GLM; carried out using SPM12, Friston et al., 2007).[6] GLM is a hierarchical model with two levels that is typically used in fMRI data analysis (Poldrack et al., 2011), and its use within neuro-computational models of language processing for continuous, naturalistic fMRI studies is well-established (Brennan et al., 2012; Brennan, 2016; Willems et al., 2016; Bhattasali et al., 2018; Li et al., 2018; Bhattasali et al., 2019). At the first level of the GLM model, the data for each subject is modeled separately to calculate subject-specific parameter estimates and within-subject variance such that for each subject, a regression model is estimated for each voxel against the fMRI time series. The second-level model takes subject-specific parameter estimates as input and uses the between-subject variance to make statistical inferences about the larger population. The end result is a time series linear regression between the estimated fMRI BOLD signal and observed BOLD signals across the whole brain. Correlations between time series can then be computed with determinations of statistical significance, with suitable corrections for multiple comparisons.

4.1 Analysis 1: ngram surprisal vs. Topical surprisal

We regressed the word-by-word predictors against fMRI timecourses recorded during passive story-listening in a whole-brain analysis. The regressors were time-locked at the offset of each word in the audiobook. For each of the 15,388 words in the story, their timestamps were estimated using Praat TextGrids (Boersma and Weenink). Each word was annotated with its 5gram surprisal and the 6,243 non-function words were annotated with its topical surprisal values, as described in §3.5.

Additionally, we entered four regressors of non-interest into the GLM analysis: word-offset, word

[5] For more details about the hemodynamic response, see chapter 2 of Kemmerer (2014).

[6] Processing time on a Mac OS 10.13 takes 1.5 hours per subject and increases linearly with additional subjects.

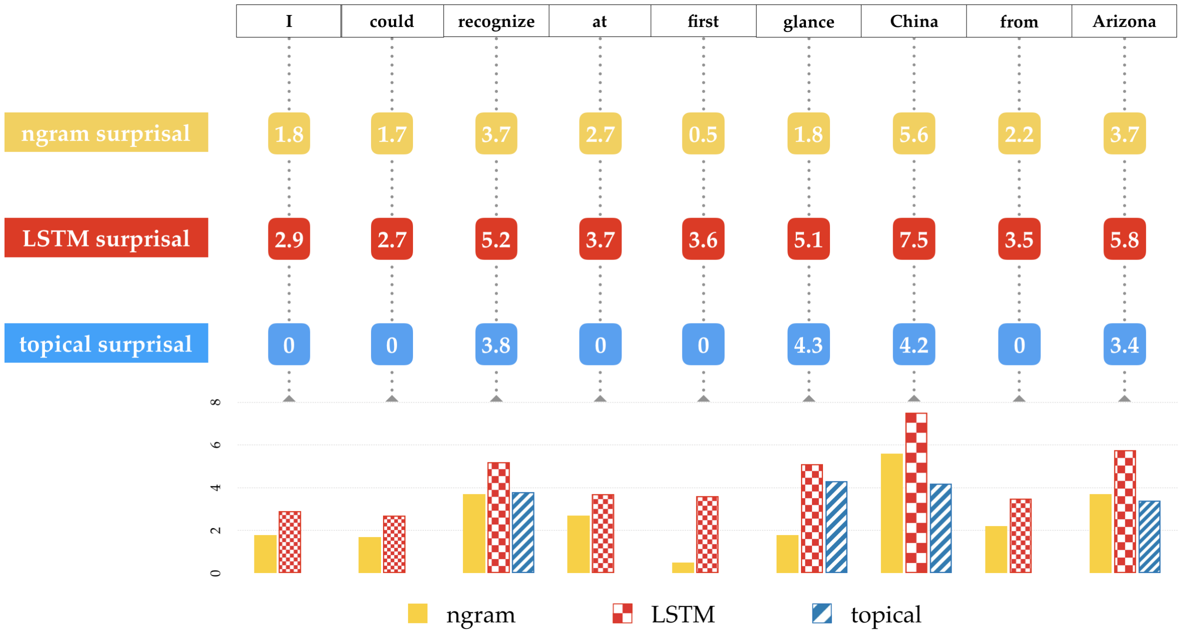


Figure 2: Comparing the word-by-word predictors on a single sentence: ngram surprisal (in yellow), neural surprisal (in red), topical surprisal (in blue). Values scaled for visualization purposes.

frequency, pitch, and intensity, which serve to improve the sensitivity, specificity and validity of activation maps (Bullmore et al., 1999; Lund et al., 2006). These were added to ensure that conclusions about lexical surprisal and topical surprisal would be specific to the cognitive processes they were taken to instantiate, as opposed to more general aspects of speech perception. These regressors were not orthogonalized.

4.2 Analysis 2: LSTM surprisal vs. Topical surprisal

Analysis 2 uses the same predictors as in Analysis 1, except that we use an LSTM language model to calculate lexical surprisal. Each word is annotated with its corresponding LSTM surprisal value (as described in §3.5), instead of 5gram surprisal, along with topical surprisal value given to the non-function words. These regressors were also not orthogonalized.

4.3 Group-level Analysis

In the second-level group analysis, each contrast was analyzed separately at the group-level. An 8 mm FWHM Gaussian smoothing kernel was applied on the contrast images from the first-level analysis to counteract inter-subject anatomical variation.

5 Results and Discussion

To begin with necessary details, behavioural results in the comprehension task confirmed that subjects were listening attentively to the auditory story presentation: across 51 participants, average accurate responses to the comprehension questions was 90% (SD = 3.7%). All whole-brain effects reported below survived a $p < 0.05$ Family-Wise-Error voxel correction for multiple comparisons which resulted in T-scores > 5.3. All brain region labels are from the Harvard-Oxford Cortical Structure Atlas.

Turning to the results of our analyses, functional localization identified using fMRI — via significant correlation with surprisal models — is interpreted to show which brain regions are recruited in processing the different types of contextual information captured by those models. To summarize, we observe a functionally distinct network that shows the difference between the influence of broad contextual cues and local contextual cues during sentence processing.

5.1 Analysis 1: Group-level results for ngram surprisal vs. topical surprisal

Whole-brain contrasts show that broad contextual cues and local contextual cues implicate different brain regions with no overlap (see Fig. 3). We observe a right-lateralized pattern of activation for topical surprisal (instantiating broad context) with

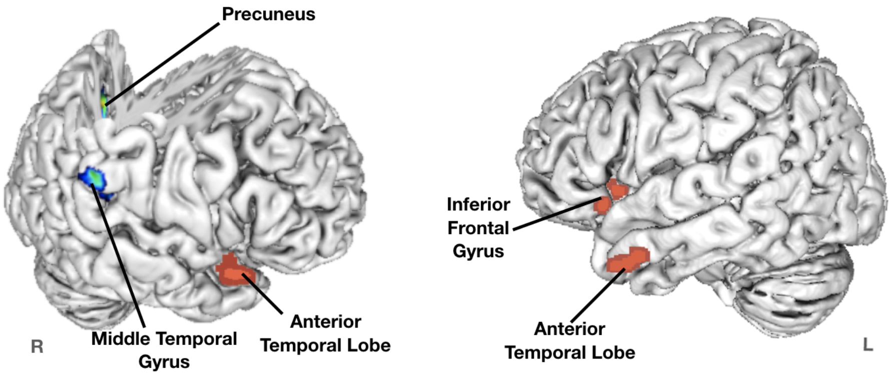


Figure 3: Whole brain contrast image with significant clusters for 5gram surprisal (in orange) and topical surprisal (in blue) after FWE voxel correction with $p < 0.05$. Table with significant clusters of peak activation included in Supplementary Materials.

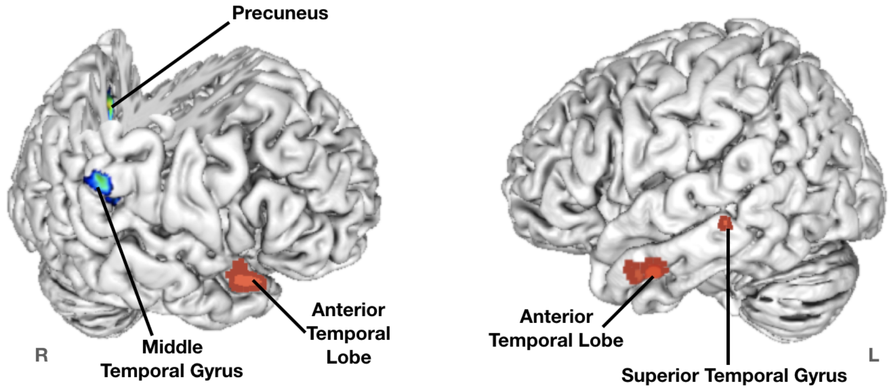


Figure 4: Whole brain contrast image with significant clusters for LSTM surprisal (in orange) and topical surprisal (in blue) after FWE voxel correction with $p < 0.05$. Table with significant clusters of peak activation are included in Supplementary Materials.

neural activation in the Precuneus and Middle Temporal Gyrus. For 5gram surprisal, the significant clusters are in the bilateral Anterior Temporal Lobe and left Inferior Frontal Gyrus.

5.2 Analysis 2: Group level results for LSTM surprisal vs. topical surprisal

Regression analyses localized the activation patterns for local and broad context to different areas (see Fig. 4). The peak activation for LSTM surprisal (instantiating local context) was observed in bilateral Anterior Temporal Lobe, along with a small cluster in left Superior Temporal Gyrus. Significant clusters for topical surprisal (instantiating broad context) were seen in the right Precuneus and right Middle Temporal Gyrus.

5.3 Discussion

In terms of Marr's (1982) levels, studies of the kind described here involve a linkage between proposals at the algorithmic-representational level and processing at the implementation level. Specifically, the logic of surprisal-based studies in computational cognitive neuroscience is based on the idea that when a word is less expected, it gives rise to increased brain activity due to increased cognitive load. Different instantiations of surprisal are used to model aspects of processing taking place during language comprehension, and correlations of surprisal with increased brain activity provide evidence about those aspects of the human comprehension process. Surprisal defined using ngrams embodies the use of sequential contextual representations during sentence processing. Syntactic surprisal embodies the use of hierarchical syntactic representations.

The present neuroimaging study introduces a new model, topical surprisal, which concerns the use of broader contextual information during processing — the topical probability of a word as defined in Eq (2) can be interpreted as an expected value of the word's probability given the topic being discussed, under the posterior topic distribution for the context. Our goal in this paper was to contrast context within a sentence with broader topical context. This can be viewed as a natural step in a progression from narrow ngram surprisal to sentential (syntactic and LSTM) surprisal to broader context that comprehenders might be taking into account.

The results in §5 based on our new model show that the patterns of activation for topical surprisal differ from those of lexical surprisal and syntactic surprisal, notably involving the right hemisphere. The centrality of the right Middle Temporal Gyrus and right Precuneus is consistent with previous studies demonstrating the role of those regions in broader contextual prediction during language comprehension: studies on narrative shifts (Whitney et al., 2009), contrasting sentences and narratives (Xu et al., 2005), contrasting coherent and incoherent narratives (Maguire et al., 1999), and sentences with and without preceding context (Raposo and Marques, 2013) that have found these same brain regions are involved in processing broader context and discourse-level information. The converging evidence confirms that our formalization of topical surprisal represents a cognitively plausible metric. Moreover, the neural correlates for topical surprisal corroborate previous work on lexical access and semantic integration (Binder et al., 2009; Graves et al., 2010; Hickok and Poeppel, 2007; Hagoort and Indefrey, 2014), suggesting that this broader contextual prediction is involved in these psycholinguistic processes beyond the sentence level.

Our novel approach to investigating contextual fit beyond the sentence level is also broadly consistent with prior results demonstrating how cortical hierarchy overlaps with language regions by using increasingly larger temporal receptive windows Lerner et al. (2011). Our results can be taken to support the argument that smaller versus larger temporal receptive windows implicate regions associated with lower-level and higher-level tasks respectively, a connection we plan to explore further.

Looking just at the ngram and LSTM models of lexical surprisal, our results provide additional corroboration for previous findings in the cognitive neuroscience literature involving sequential and syntactic processing (Willems et al., 2016; Brennan, 2016; Lopopolo et al., 2017; Shain et al., 2020). They also constitute an addition to the literature on understanding the linguistic properties captured by deep learning architectures. Numerous authors have shown that LSTM models capture not only sequential but also longer-distance structural dependencies (Linzen et al., 2016; Gulordava et al., 2018; van Schijndel and Linzen, 2018; Kuncoro et al., 2018; Futrell et al., 2019). In our study, we find that, while overlapping in the bilateral Anterior Temporal Lobe, the ngram and syntactic surprisal models also give rise to differently localized brain activity: the ngram model implicates the left Inferior Frontal Gyrus (IFG), while the LSTM surprisal model implicates the left Superior Temporal Gyrus (STG). The key observation here is that left-lateralized STG activity is uncontroversially correlated with syntactic processing (Pallier et al., 2011; Thompson and Meltzer-Asscher, 2014; Bhattasali et al., 2019; Shain et al., 2020). In contrast, the functional role of left IFG has variously been interpreted as involving combinatorial, sequential, or syntactic processes (Sahin et al., 2009; Snijders et al., 2009; Pallier et al., 2011; Brennan et al., 2016). The patterns of activity in this study therefore provide support from the implementation level for the idea that LSTMs are capturing aspects of syntactic representation that ngram models do not. Narrowing down the nature of those differences (e.g., sequential versus structural, or short- versus long-distance syntactic dependencies) remains a subject for future work.

6 Conclusion

The present study posed the questions of how broader topical context influences expectations in human sentence comprehension, and how local versus broader contexts might be processed differently in the brain. To address those questions we have introduced topical surprisal, a straightforward and intuitive extension to the highly productive surprisal-based paradigm in computational psycho- and neurolinguistics that employs topic modeling to estimate word probabilities conditioned on contexts beyond the sentence level.

Using analysis of fMRI brain imaging during naturalistic listening, we showed that the processing of broader topical context gives rise to neural activity in different brain regions than local contextual prediction as modeled using ngrams or an LSTM. The brain regions we identified turn out to be consistent with prior studies looking at neural correlates for processing of narratives and discourse.

In addition, we explored the neuro-anatomical

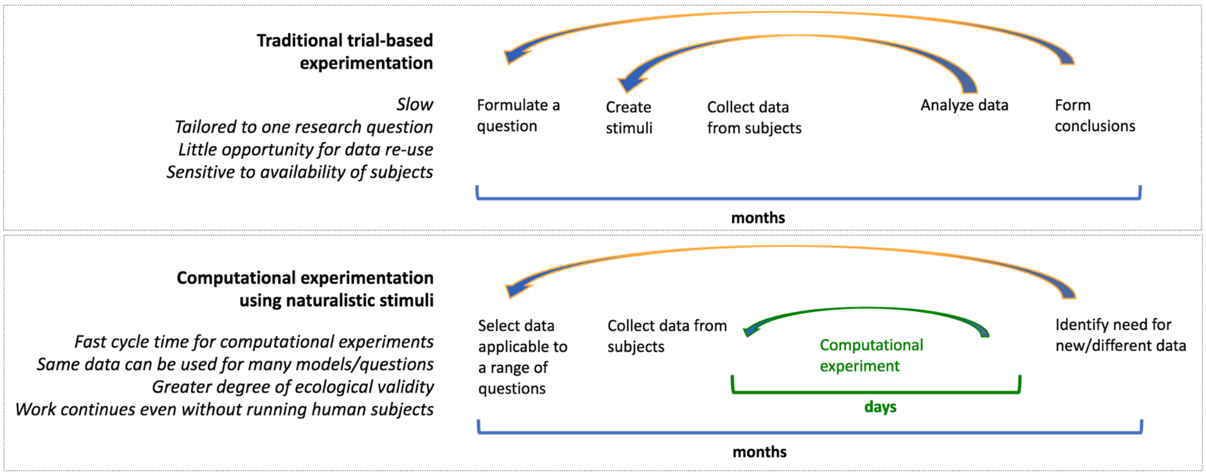

Traditional trial-based experimentation
Slow
Tailored to one research question
Little opportunity for data re-use
Sensitive to availability of subjects
Formulate a question
Create stimuli
Collect data from subjects
Analyze data
Form conclusions
months
Computational experimentation using naturalistic stimuli
Fast cycle time for computational experiments
Same data can be used for many models/questions
Greater degree of ecological validity
Work continues even without running human subjects
Select data applicable to a range of questions
Collect data from subjects
Computational experiment
days
Identify need for new/different data
months

Figure 5: Visual comparison between two different experimental paradigms for cognitive neuroscience research

correlates of ngram and LSTM processing and obtained results that are consistent with claims in the deep learning literature regarding the sensitivity of LSTMs to long-distance syntactic structure.

More generally, this paper adds another data point demonstrating the relevance of tools from computational linguistics in cognitive neuroscience research (Brennan and Hale, 2019; Jain and Huth, 2018; Toneva et al., 2020) and the value of naturalistic stimuli in contextually situated and ecologically valid research (Maguire, 2012; Brennan, 2016; Hamilton and Huth, 2020).

Finally, we note that the paradigm we have employed here — computational modeling with previously collected natural-listening data — promotes reusability of datasets and replicability of results, and safeguards against unexpected delays in data collection such as a pandemic. Even more important, it offers a rapid experimental cycle dramatically better suited to computational research than traditional, trial-based methods in psycholinguistic and neurolinguistic research (Figure 5). As such, computational experimentation with naturalistic stimuli presents an invitation to computational linguists to collaborate with cognitive neuroscientists and apply their skills in operationalizing and testing hypotheses about neurocognitive mechanisms in sentence processing.

Ethical Considerations

This scientific study was reviewed and approved by the Cornell University Institutional Review Board, and human subject participation was conducted according to standard practices for research of this kind, including compensation of $65 for participants who completed the study. No personally identifiable data are stored. As is frequently the case for academic human subjects studies, our sample is drawn from a university population and may not be representative of the population at large. In addition, safety protocols required excluding participants with metal in their body (e.g. surgical implants, fresh tattoos), and following standard practice we included only participants with no history of psychiatric, neurological, or other medical illness or history of drug or alcohol abuse that might compromise cognitive functions, nor anyone taking antidepressant or psychoactive medications. Although the study was conducted in English, it included international and bilingual participants.

Acknowledgments

We would like to thank John Hale for his help in designing the study. We would also like to thank the audiences at AMLaP 2020, SNL 2020, the UMD Cognitive Neuroscience of Language group, and the MIT Computational Psycholinguistics group for their feedback on earlier stages of this study. We are also grateful to the reviewers for their thoughtful comments.

This material is based upon work supported by ONR MURI Award N00014–18–1–2670. The data was collected under NSF Grant No. 1607441.

References

Shohini Bhattasali, Murielle Fabre, and John Hale. 2018. Processing MWEs: Neurocognitive bases of verbal MWEs and lexical cohesiveness within MWEs. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and*

Constructions (LAW-MWE-CxG-2018), pages 6–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shohini Bhattasali, Murielle Fabre, Wen-Ming Luh, Hazem Al Saied, Mathieu Constant, Christophe Pallier, Jonathan R Brennan, R Nathan Spreng, and John Hale. 2019. Localising memory retrieval and syntactic composition: An fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, 34(4):491–510.

Jeffrey R. Binder, Rutvik H. Desai, William W. Graves, and Lisa L. Conant. 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–2796.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *the Journal of machine Learning research*, 3:993–1022.

Paul Boersma and David Weenink. *Praat: Doing phonetics by computer*.

Jonathan Brennan. 2016. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.

Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J Heeger, and Liina Pylkkänen. 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and language*, 120(2):163–173.

Jonathan R. Brennan and John T Hale. 2019. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1).

Jonathan R. Brennan, Edward P. Stabler, Sarah E. Van Wagenen, Wen-Ming Luh, and John T. Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94.

Christian Brodbeck, L. Elliot Hong, and Jonathan Z. Simon. 2018. Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, 28(24):3976–3983.

Edward T. Bullmore, Michael J. Brammer, Sophia Rabe-Hesketh, Vivienne A. Curtis, Robin G. Morris, Steve C.R. Williams, Tonmoy Sharma, and Philip K. McGuire. 1999. Methods for diagnosis and treatment of stimulus-correlated motion in generic brain activation studies using fMRI. *Human brain mapping*, 7(1):38–48.

Mark Davies. 2008. *The Corpus of Contemporary American English (COCA): 560 million words, 1990–present*.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Allyson Ettinger, Naomi Feldman, Philip Resnik, and Colin Phillips. 2016. Modeling N400 amplitude using vector space models of word representation. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pages 1445–1450.

K.J. Friston, J. Ashburner, S.J. Kiebel, T.E. Nichols, and W.D. Penny, editors. 2007. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

William W. Graves, Jeffrey R. Binder, Rutvik H. Desai, Lisa L. Conant, and Mark S. Seidenberg. 2010. Neural correlates of implicit and explicit combinatorial semantic processing. *Neuroimage*, 53(2):638–646.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 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 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Peter Hagoort and Peter Indefrey. 2014. The neurobiology of language beyond single words. *Annual review of neuroscience*, 37:347–362.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics.

Liberty S. Hamilton and Alexander G. Huth. 2020. The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696.

John M. Henderson, Wonil Choi, Matthew W. Lowder, and Fernanda Ferreira. 2016. Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *Neuroimage*, 132:293–300.

Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Shailee Jain and Alexander G. Huth. 2018. Incorporating context into language encoding models for fMRI. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6629–6638.

Katerina D. Kandylaki and Ina Bornkessel-Schlesewsky. 2019. From story comprehension to the neurobiology of language. *Language, Cognition and Neuroscience*, 34(4):405–410.

David Kemmerer. 2014. *Cognitive neuroscience of language*. Psychology Press.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Melbourne, Australia. Association for Computational Linguistics.

Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.

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

Jixing Li, Murielle Fabre, Wen-Ming Luh, and John Hale. 2018. The role of syntax during pronoun resolution: Evidence from fMRI. In *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, pages 56–64, Melbourne. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Alessandro Lopopolo, Stefan L. Frank, Antal Van den Bosch, and Roel M. Willems. 2017. Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PloS one*, 12(5):e0177794.

Torben E. Lund, Kristoffer H. Madsen, Karam Sidaros, Wen-Lin Luo, and Thomas E. Nichols. 2006. Non-white noise in fMRI: does modelling have an impact? *Neuroimage*, 29(1):54–66.

Eleanor A Maguire. 2012. Studying the freely-behaving brain with fMRI. *Neuroimage*, 62(2):1170–1176.

Eleanor A. Maguire, Christopher D. Frith, and R. G. M. Morris. 1999. The functional neuroanatomy of comprehension and memory: The importance of prior knowledge. *Brain*, 122(10):1839–1850.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit.

James Michaelov and Benjamin Bergen. 2020. How well does surprisal explain n400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663, Online. Association for Computational Linguistics.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

Richard C. Oldfield. 1971. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9(1):97–113.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527.

Jonathan W. Peirce. 2007. PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1):8–13.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Russell A. Poldrack, Jeanette A. Mumford, and Thomas E. Nichols. 2011. *Handbook of functional MRI data analysis*. Cambridge University Press.

Ana Raposo and J. Frederico Marques. 2013. The contribution of fronto-parietal regions to sentence comprehension: Insights from the Moses illusion. *NeuroImage*, 83:431–437.

Radim Řehůřek and Petr Sojka. Gensim – Statistical semantics in Python.

Ned T. Sahin, Steven Pinker, Sydney S. Cash, Donald Schomer, and Eric Halgren. 2009. Sequential processing of lexical, grammatical, and phonological information within Broca's area. *Science*, 326(5951):445–449.

Marten van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710, Brussels, Belgium. Association for Computational Linguistics. `https://github.com/vansky/neural-complexity`.

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307.

Tineke M. Snijders, Theo Vosse, Gerard Kempen, Jos J.A. Van Berkum, Karl Magnus Petersson, and Peter Hagoort. 2009. Retrieval and unification of syntactic structure in sentence comprehension: An fMRI study using word-category ambiguity. *Cerebral cortex*, 19(7):1493–1503.

Miloš Stanojević, Shohini Bhattasali, Donald Dunagan, Luca Campanelli, Mark Steedman, Jonathan Brennan, and John Hale. 2021. Modeling incremental language comprehension in the brain with Combinatory Categorial Grammar. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 23–38, Online. Association for Computational Linguistics.

Cynthia K Thompson and Aya Meltzer-Asscher. 2014. Neurocognitive mechanisms of verb argument structure processing. *Structuring the argument*, pages 141–168.

Mariya Toneva, Tom M. Mitchell, and Leila Wehbe. 2020. Combining computational controls with natural text reveals new aspects of meaning composition. *bioRxiv*.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pages 14928–14938.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar. Association for Computational Linguistics.

Carin Whitney, Walter Huber, Juliane Klann, Susanne Weis, Sören Krach, and Tilo Kircher. 2009. Neural correlates of narrative shifts during auditory story comprehension. *Neuroimage*, 47(1):360–366.

Roel M. Willems. 2015. Introduction. *Cognitive neuroscience of natural language use*, pages 1–7.

Roel M. Willems, Stefan L. Frank, Annabel D. Nijhof, Peter Hagoort, and Antal Van den Bosch. 2016. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.

Jiang Xu, Stefan Kemeny, Grace Park, Carol Frattali, and Allen Braun. 2005. Language in context: Emergent features of word, sentence, and narrative comprehension. *Neuroimage*, 25(3):1002–1015.

Shulin Zhang. 2020. Human brain networks for semantic roles. Master's thesis, University of Georgia.

Appendix for Using surprisal and fMRI to map the neural bases of broad and local contextual prediction during natural language comprehension

1 Data Collection

Imaging was performed using a 3T MRI scanner (Discovery MR750, GE Healthcare, Milwaukee, WI) with a 32-channel head coil at the Cornell MRI Facility. Blood Oxygen Level Dependent (BOLD) signals were collected using a T2 -weighted echo planar imaging (EPI) sequence (repetition time: 2000 ms, echo time: 27 ms, flip angle: 77deg, image acceleration: 2X, field of view: 216 x 216 mm, matrix size 72 x 72, and 44 oblique slices, yielding 3 mm isotropic voxels). Anatomical images were collected with a high resolution T1-weighted (1 x 1 x 1 mm^3 voxel) with a Magnetization-Prepared RApid Gradient-Echo (MP-RAGE) pulse sequence.

2 Preprocessing

Primary preprocessing steps were carried out in AFNI version 16 (Cox, 1996) and include motion correction, coregistration, and normalization to standard MNI space. After the previous steps were completed, ME-ICA (Kundu et al., 2012) was used to further preprocess the data. ME-ICA is a denoising method which uses Independent Components Analysis to split the T2*-signal into BOLD and non-BOLD components. Removing the non-BOLD components mitigates noise due to motion, physiology, and scanner artifacts (Kundu et al., 2017).

3 Correlation Matrix for Predictors

Fig. 1 shows the correlation matrix for the three surprisal predictors.

4 Group-level results

Table 1 shows the significant clusters of activation for topical surprisal using brain region labels from the Harvard-Oxford Cortical Structure Atlas.

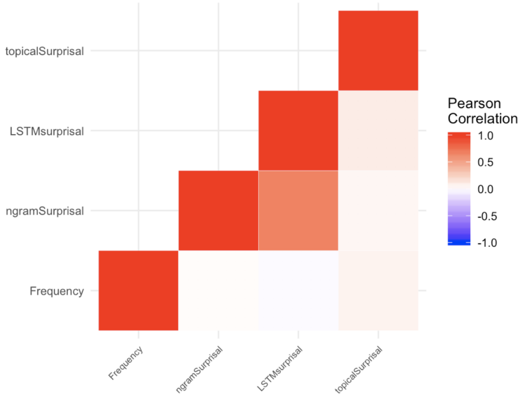


Figure 1: Correlation matrix (Pearson's r) of the convolved regressors included in the GLM models reported in Analysis 1 and Analysis 2.

Regions	Cluster size	MNI Coordinates			p-value	T-score
	(in voxels)	x	y	z	(corrected)	(peak-level)
R Middle Temporal Gyrus	497	50	-50	18	0.000	8.46
R Precuenus	253	10	-62	30	0.011	5.83

Table 1: Significant clusters for topical surprisal after FWE voxel correction with $p < 0.05$. Peak activation is given in MNI coordinates and p-values are reported at peak-level after voxel correction.

References

Robert W. Cox. 1996. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173.

Prantik Kundu, Souheil J. Inati, Jennifer W. Evans, Wen-Ming Luh, and Peter A. Bandettini. 2012. Differentiating bold and non-bold signals in fMRI time series using multi-echo epi. *Neuroimage*, 60(3):1759–1770.

Prantik Kundu, Valerie Voon, Priti Balchandani, Michael V. Lombardo, Benedikt A. Poser, and Peter A. Bandettini. 2017. Multi-echo fMRI: A review of applications in fMRI denoising and analysis of bold signals. *NeuroImage*, 154:59 – 80.