bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The meaning that emerges from combining words is robustly localizable in space but not in time

Mariya Toneva[1,2], Tom M. Mitchell[1,2], and Leila Wehbe[*1,2]

[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA
[2]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, USA

Abstract

To study a core component of human intelligence—our ability to combine the meaning of words—neuroscientists look for neural *correlates* of meaning composition, such as brain activity proportional to the difficulty of understanding a sentence. However, little is known about the *product* of meaning composition in the brain—the combined meaning of words beyond their individual meaning. We term this product "supra-word meaning" and devise a computational representation for it by using recent neural network algorithms and a new technique to disentangle composed- from individual-word meaning. Using functional magnetic resonance imaging, we find that supra-word meaning is processed in the bilateral anterior and posterior temporal lobes. Surprisingly, we cannot detect supra-word meaning in magnetoencephalography. These results suggest that composed meaning is maintained through a different neural mechanism that does not consist of synchronized cell firing. This difference in sensitivity has implications for past neuroimaging results and future neuroimaging studies.

Introduction

Understanding language in the real-world requires us to compose the meaning of individual words in a way that makes the final composed product more meaningful than the string of isolated words. For example, we understand the statement that "Mary finished the apple" to mean that Mary finished *eating* the apple, even though "eating" is not explicitly specified (Pylkkänen, 2020). This *supra-word meaning*, or the product of meaning composition beyond the meaning of individual words, is at the core of language comprehension, and its neurobiological bases and processing mechanisms must be specified in the pursuit of a complete theory of language processing in the brain.

However, neuroscientists have instead investigated *correlates* of meaning composition that may fail to capture or isolate the supra-word meaning. In one line of research, neuroscientists follow classical neuroimaging approaches that consist of contrasting a condition of interest (e.g., semantic surprise, full sentences or erroneous sentences) with a control condition (e.g.,

[*]Correspondence: lwehbe@cmu.edu.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

no surprise, disconnected words or correct sentences). For example, they observe differences in brain recordings when processing an unexpected versus an expected word in a specific context (Kutas and Federmeier, 2011; Kuperberg *et al.*, 2003; Kuperberg, 2007) or a monotonic increase in neural activity over the course of reading a sentence (Fedorenko *et al.*, 2016). Even though such studies have been pivotal in beginning to study the processes behind meaning composition, we argue that their findings are related to the process of integrating supra-word meaning, while missing other key components, such as the storage and maintenance of the current supra-word meaning. In a different line of work, neuroscientists build computational models of meaning through Natural Language Processing (NLP) embeddings of words and sentences (Mitchell *et al.*, 2008; Sudre *et al.*, 2012; Wehbe *et al.*, 2014b,a; Huth *et al.*, 2016; Jain and Huth, 2018; Toneva and Wehbe, 2019; Fyshe *et al.*, 2019). Thanks to these studies, we are starting to uncover some properties of meaning representation, such as the fact that neural activity associated with single word meaning is distributed (Mitchell *et al.*, 2008; Huth *et al.*, 2016). However, the neural substrates of composed meaning and the mechanism by which it is represented are still elusive and we are far from converging on a mechanistic, algorithmic understanding of meaning composition beyond individual words.

In this work, we study the brain representation of supra-word meaning by using data from naturalistic reading and two neuroimaging modalities. More formally, we define "supra-word meaning" as the composed meaning of a sequence of words that is not part of the corresponding bag-of-words, i.e., it is the new meaning formed by combining a sequence of words that is not included in the isolated meaning of those words. In addition to implied meaning, other examples of supra-word meaning include: 1) a specific contextualized meaning of a word or phrase (e.g. "green banana" evokes the meaning of an unripe, rather than simply green-colored, banana) that can also distinguish between different senses of the same word (e.g. "play a game" versus "theater play"), and 2) the different meaning of two events that can be described with the same words but reversed semantic roles (e.g. "John gives Mary an apple" and "Mary gives John an apple").

We create a computer representation for this supra-word meaning, derived from recently developed natural language processing algorithms (Peters *et al.*, 2018). We find that this representation of supra-word meaning predicts fMRI activity in the anterior and posterior temporal cortices, suggesting that these areas represent composed meaning. These areas can be sub-divided into clusters that appear to process different parts of composed meaning. We also find that it is very hard to detect the representation of composed meaning from MEG activity. MEG has been shown to reveal signatures of the *computations* involved in incorporating a word into a sentence (Halgren *et al.*, 2002; Lyu *et al.*, 2019), which are themselves a function of the composed meaning of the words seen so far. However, our results suggest that the sustained *representation* of the composed meaning may rely on neural mechanisms that do not lead to reliable MEG activity. This novel finding calls for a retrofitted understanding of the body of literature on meaning composition and has important implications for the future of brain-computer interfaces.

Model

We built on recent progress in NLP that has resulted in algorithms that can capture the meaning of words in a particular context. One such algorithm is ELMo (Peters *et al.*, 2018), a

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

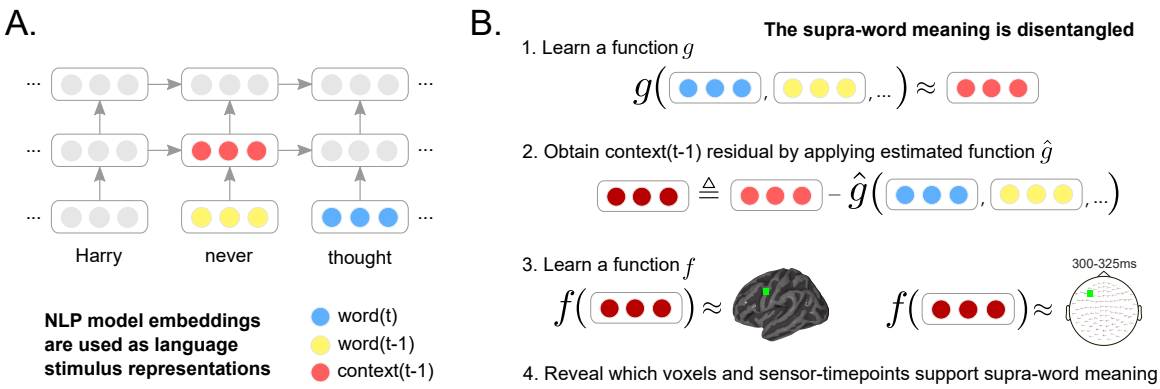


Figure 1: Approach. **(A)** Stimulus representations are obtained from an NLP model that has captured language statistics from millions of documents. This model represents words using context-free embeddings (shown in yellow and blue) and context embeddings (shown in red). Context embeddings are obtained by continuously integrating each new word's context-free embedding with the most recent context embedding. **(B)** Supra-word meaning is modeled by obtaining the residual information in the context embeddings after removing information related to the word embeddings. The supra-word meaning is used as an input to an encoding model f, revealing which fMRI voxels and MEG sensor-timepoints are modulated by supra-word meaning.

powerful language model with a bi-directional Long Short-Term Memory (LSTM) architecture. ELMo estimates a *contextualized* embedding for a word by combining a *non-contextualized* fixed input vector for that word with the internal state of a forward LSTM (containing information from previous words) and a backward LSTM (containing information from future words). To capture information about word t, we used the input vector for word t. To capture information about the context preceding word t, we used the internal state of the forward LSTM computed at word $t-1$ (Fig. 1A). We did not include information from the backward LSTM, since it contains future words which have not yet been seen at time t.

To study supra-word meaning, the meaning that results from the composition of words should be isolated from the individual word meaning. ELMo's context embeddings contain information about individual words (e.g., 'finished', 'the', and 'apple' in the context 'finished the apple') in addition to the implied supra-word meaning (e.g., eating) (Appendix 1, Suppl. Fig. 1). We post-processed the context embeddings produced by ELMo to remove the contribution due to the context-independent meanings of individual words. We constructed a "residual context embedding" by removing the shared information between the context embedding and the meanings of the individual words (Fig. 1B).

To investigate the neural substrates and temporal dynamics of supra-word meaning, we trained encoding models, as a function of supra-word meaning, to predict the brain recordings of nine fMRI participants and eight MEG participants as they read a chapter of a popular book in rapid serial visual presentation. The encoding models predict each fMRI voxel and MEG sensor-timepoint, from the text read by the participant up to that time point. The prediction performance of these models was tested by computing the correlation between the model predictions and the true held-out brain recordings. Hypothesis tests were used

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to identify fMRI voxels and MEG sensor-timepoints that were significantly predicted by supra-word meaning. For more details about the training procedure and hypothesis tests, see Materials and Methods.

Results

Detecting regions that are predicted by supra-word meaning

To identify brain areas that represent supra-word meaning, we focus on the fMRI portion of the experiment. We find that many areas previously implicated in language-specific processing (Fedorenko *et al.*, 2010; Fedorenko and Thompson-Schill, 2014) and word semantics (Binder *et al.*, 2009) are significantly predicted by the full context embeddings across subjects (voxel-level permutation test, Benjamini-Hochberg FDR control at 0.01 (Benjamini and Hochberg, 1995)). These areas include the bilateral posterior and anterior temporal cortices, angular gyri, inferior frontal gyri, posterior cingulate, and dorsomedial prefrontal cortex (Fig. 2A). A subset of these areas is also significantly predicted by residual context embeddings. To quantify these observations, we select regions of interest (ROIs) based on the works above (Fedorenko *et al.*, 2010; Binder *et al.*, 2009), using ROI masks that are entirely independent of our analyses and data (see Materials and Methods). Full context embeddings predict a significant proportion of the voxels within each ROI across all 9 participants (Fig. 2B; ROI-level Wilcoxon signed-rank test, $p < 0.05$, Holm-Bonferroni correction (Holm, 1979)). In contrast, residual context embeddings predict a significant proportion of only the anterior and posterior temporal lobes. While the full context embedding is predictive of much of the fMRI recordings across the brain, the supra-word meaning is selectively predictive of two language regions - the anterior (ATL) and posterior temporal lobes (PTL).

Do the parts of the ATL and PTL that are predicted by supra-word meaning process the same information? Inspired by temporal generalization matrices (King and Dehaene, 2014), we introduce *spatial generalization matrices* that estimate the pairwise similarity of voxel representations (see Materials and Methods). The spatial generalization matrices reveal that the PTL can be divided into two main clusters such that the models of voxels in one cluster can also predict other voxels in that cluster but not in the other cluster (Fig. 2C and Suppl. Fig. 3; voxel-level permutation test, Benjamini-Hochberg FDR controlled at level 0.01). Furthermore, the models of voxels within one of the PTL clusters, but not the other, significantly predict voxels in the ATL. The division of the PTL into two clusters, one of which is predictive of the ATL, can be observed within- (Fig. 2C, left), and across-participants (Fig. 2C, right). In contrast, the ATL voxels show only one cluster of voxels that are predictive both of other ATL voxels and also of PTL voxels (Suppl. Fig. 2). This pattern indicates that the organization of information in the ATL and parts of the PTL is shared and consistent across participants. To localize this shared representation, we visualize how well each ATL and PTL voxel predicts the other participants' ATLs (Fig. 2D and Suppl. Fig. 4). ATL voxels are predictive of significant proportions of the ATL across participants, reinforcing the single cluster of ATL voxels observed in the spatial generalization matrices. Much of the left PTL predicts a significant proportion of the ATL across participants, whereas much of the right PTL does not (ROI-level Wilcoxon signed-rank test, $p < 0.05$, Holm-Bonferroni correction). The left PTL appears further subdivided, with a cluster of voxels in the posterior

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

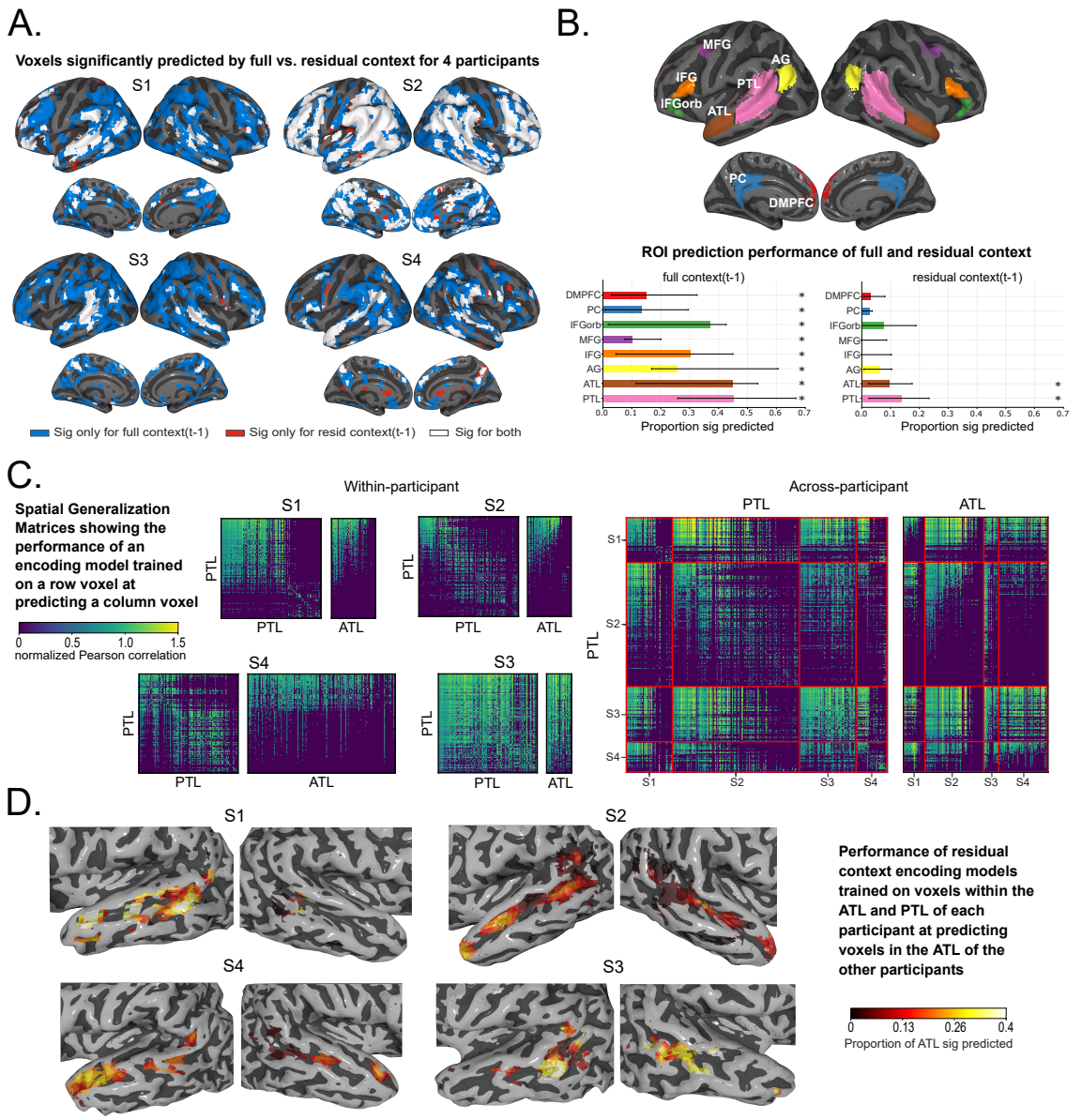


Figure 2: fMRI results. Voxel-level significance is FDR corrected at $\alpha = 0.01$. **(A)** Voxels significantly predicted by full-context embeddings (blue), residual-context embeddings (red), or both (white), visualized in MNI space. Most of the temporal cortex and IFG is predicted by full context embeddings, with residual context embeddings mostly predicting a subset of those areas. **(B)** ROI-level results. (Top) Language system ROIs (Fedorenko *et al.*, 2010) and two semantic ROIs (Binder *et al.*, 2009). (Bottom) Proportion of ROI voxels significantly predicted by (Left) full context and (Right) residual context embeddings. Displayed are the median proportions across all participants and the medians' 95% confidence intervals. Full context predicts all ROIs (ROI-level Holm-Bonferroni correction, $p < 0.05$), while residual context predicts only bilateral ATL and PTL. **(C)** Spatial Generalization Matrices. Models trained to predict PTL voxels are used to predict PTL and ATL voxels (within-participant (Left), and across-participants (Right)). PTL cross-voxel correlations form two clusters: models that predict activity for voxels in one cluster can also predict activities of other voxels in the same cluster, but not activities for voxels in the other cluster. Across participants, only one of these clusters has voxels that predict ATL voxels. **(D)** Performance of models trained on ATL and PTL voxels at predicting other participants' ATL. All participants show a cluster of predictive voxels in the pSTS.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Superior Temporal Sulcus (pSTS) being more predictive. This suggests that the ATL and the left pSTS process a similar facet of supra-word meaning.

The processing of supra-word meaning is invisible in MEG

To study the temporal dynamics of the emergence and representation of supra-word meaning, we turn to the MEG portion of the experiment (Fig. 3). We computed the proportion of sensors that are significantly predicted at different spatial granularity – the whole brain (Fig. 3A), by lobe subdivisions (Fig. 3B), and finally at each sensor neighborhood location (Fig. 3C; sensor-timepoint level permutation test, Benjamini-Hochberg FDR control at $\alpha = 0.01$). The full context embedding is significantly predictive of the recordings across all lobes (Fig. 3B, performance visualized in lighter colors; timepoint-level Wilcoxon signed-rank test, $p < 0.05$, Benjamini-Hochberg FDR correction). Surprisingly, we find that the residual context does not significantly predict any timepoint in the MEG recordings at any spatial granularity. This surprising finding leads to two conclusions. First, supra-word meaning is invisible in MEG. Second, what is instead salient in MEG recordings is information that is shared between the context and the individual words.

To understand the source of this salience, we investigated the relationship between the MEG recordings and the word embeddings for the currently-read and previously-read words. One approach to reveal this relationship is to train an encoding model as a function of the word embedding (Jain and Huth, 2018; Toneva and Wehbe, 2019). However, the word embedding corresponding to a word at position t is correlated with the surrounding word embeddings (Appendix 1, Suppl. Fig. 1). Therefore, part of the prediction performance of the word t embedding may be due to processing related to previous words. To isolate processing that is exclusively related to an individual word, we constructed "residual word embeddings", following the approach of constructing the residual context embeddings (see Materials and Methods). We observe that the residual word embeddings for the current and previous words lead to significantly worse predictions of the MEG recordings, when compared to their corresponding full embeddings (Fig. 3B, middle and right panels; timepoint-level Wilcoxon signed-rank test, $p < 0.05$, Benjamini-Hochberg FDR correction). This indicates that a significant proportion of the activity predicted by the current and previous word embeddings is due to the shared information with surrounding word embeddings. Nonetheless, we find that the residual current word embedding is still significantly predictive of brain activity everywhere the full embeddings was predictive. This indicates that properties unique to the current word are well predictive of MEG recordings at all spatial granularity. The residual previous word embedding predicts fewer time windows significantly, particularly 350-500ms post word t onset. This indicates that the activity in the first 350ms when a word is on the screen is predicted by properties that are unique to the previous word. Taken together, these results suggest that the properties of recent words are the elements that are predictive of MEG recordings, and that MEG recordings do not reflect the supra-word meaning beyond these recent words.

Lastly, we directly compared how well each imaging modality can be predicted by each meaning embedding (Fig. 4). Residual embeddings predict fMRI and MEG with significantly different accuracy (Fig. 4A), with fMRI being significantly better predicted than MEG by the residual context, and MEG being significantly better predicted by the residual of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

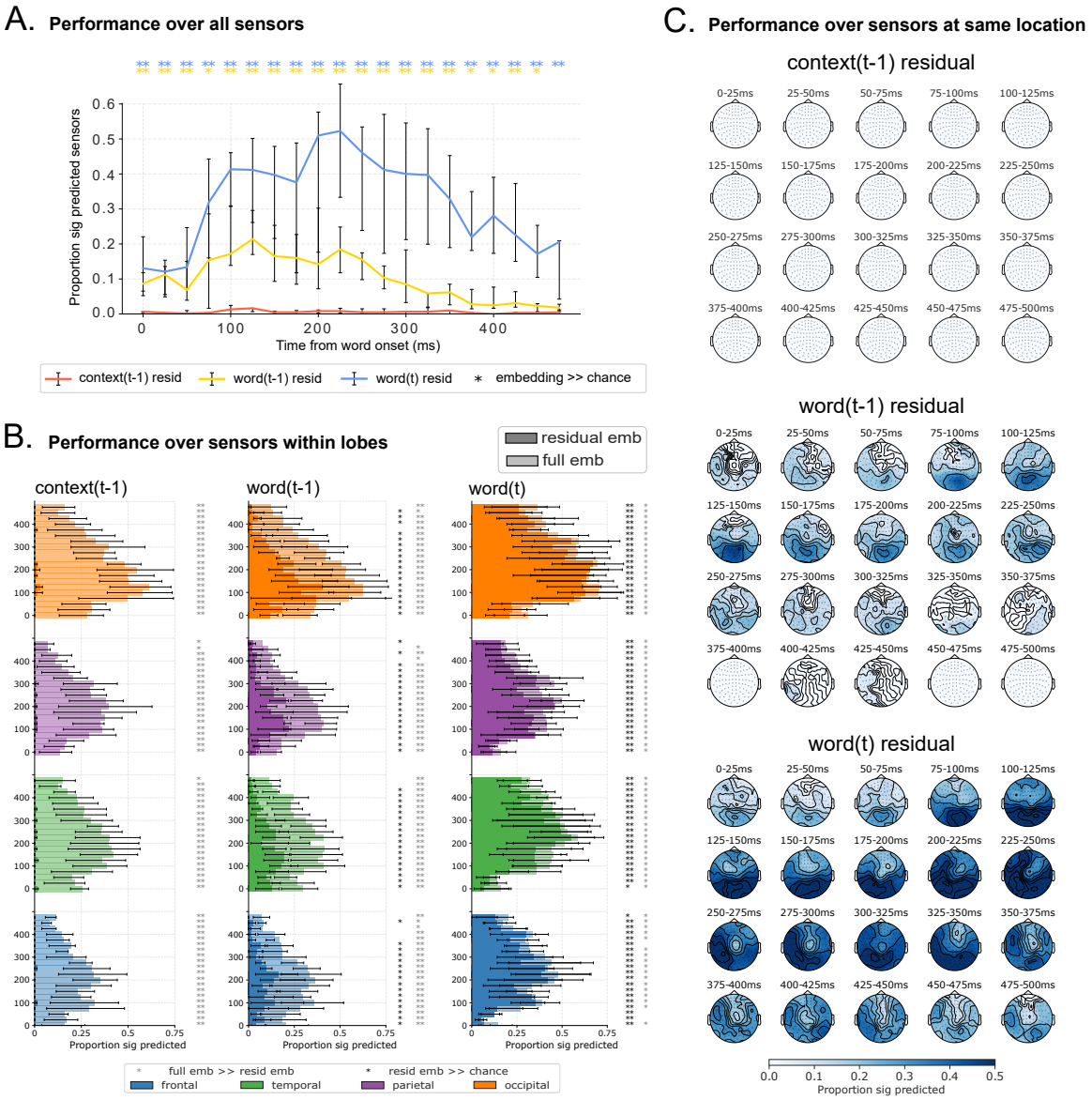


Figure 3: MEG results at different spatial granularity. All subplots present the median across participants and errorbars signify the medians' 95% confidence intervals. **(A)** Proportion of sensors for each timepoint significantly predicted by the residual embeddings. The current and previous word residuals predict activity during the entire period the word is presented (FDR corrected, $p < 0.05$), while the context residual is not predictive. **(B)** Proportion of significantly predicted sensors for each timepoint, divided by lobe. Residual embeddings performance is compared with that of full embeddings (darker and lighter colors respectively, FDR corrected, $p < 0.05$). Removing the shared information among the full current word, the previous word and the context embeddings results in a significant decrease in performance for all embeddings and lobes. The decrease in performance for the context embedding (left column) is the most drastic, with no timewindows being significant for the residual context embedding across lobes. **(C)** Proportions of sensor neighborhoods significantly predicted by each residual embedding. Only the significant proportions are displayed (FDR corrected, $p < 0.05$). Context-residuals do not predict any sensor-timepoint neighborhood while both the previous and the current word residuals predict a large subset of sensor-timepoints, with performance peaks in occipital and temporal lobes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

previous and current words (Wilcoxon rank-sum test, $p < 0.05$, Holm-Bonferroni correction). In contrast, the full context embeddings do not show a significant difference in predicting fMRI and MEG recordings(Fig. 4B). We further observe that the residual embeddings lead to an opposite pattern of prediction in the two modalities(Fig. 4C). While the residual context predicts fMRI the best out of the three residual embeddings, it performs the worst out of the three at predicting MEG (Wilcoxon signed-rank test, $p < 0.05$, Holm-Bonferroni correction). In contrast, the full context and previous word embeddings do not show a significant difference in MEG prediction(Fig. 4D), suggesting that it is the removal of individual word information from the context embedding that leads to a significantly worse MEG prediction. These findings further suggest that fMRI and MEG reflect different aspects of language processing – while MEG recordings reflect processing related to the recent context, fMRI recordings capture the contextual meaning that is beyond the meaning of individual words.

Discussion

We investigated the spatial and temporal signatures of emergent multi-word meaning, or *supra-word meaning*, by constructing a computational representation of this meaning and evaluating its ability to predict specific locations and timepoints of recorded brain activity via fMRI and MEG respectively.

We found that our devised supra-word meaning representation predicts fMRI recordings in the bilateral anterior and posterior temporal lobes (ATL and PTL). This finding supports some current hypotheses of language composition in the literature. Specifically, our results provide new evidence that the ATL processes composed meaning beyond simple concrete concepts, which supports the hypothesis that the ATL is a semantic integration hub (Visser *et al.*, 2010; Pallier *et al.*, 2011; Pylkkänen, 2020). Our results may also align with the hypothesis that the posterior superior temporal sulcus (pSTS, part of the PTL) is involved in building a type of supra-word meaning, by integrating information about the verb and its arguments with other syntactic information (Friederici, 2011; Frankland and Greene, 2015; Skeide and Friederici, 2016). Further, our findings pose questions for the theory that posits left PTL as primarily a site of lexical (i.e. word-level) semantics, and left IFG as a hub of integrated contextual information (Hagoort, 2020). Our results suggest that the role of PTL extends beyond word-level semantics. Further, we do not find evidence for supra-word meaning in left IFG, though this does not prove that left IFG does not represent supra-word meaning – the lack of significance may be due to low statistical power.

Strikingly, we found that, even though our devised supra-meaning representation predicted a significant proportion of fMRI voxels, it did not significantly predict any sensor-timepoints in MEG. Instead, the MEG recordings were significantly predicted by information unique to both the currently-read and previously-read words. These findings suggest a difference in the underlying brain processes that fMRI and MEG capture. Indeed, while it is widely known that fMRI and MEG recordings result from different physiological signals, whether they capture the same underlying brain processes is still debated (Hall *et al.*, 2014). Our results suggest that fMRI recordings are sensitive to supra-word meaning, while MEG recordings reflect instantaneous processes related to both the current word being read and the previously-read word. A likely candidate for the instantaneous process reflected in MEG is the process of integrating the current word with the previous context. The sensitivity to the previously-read

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

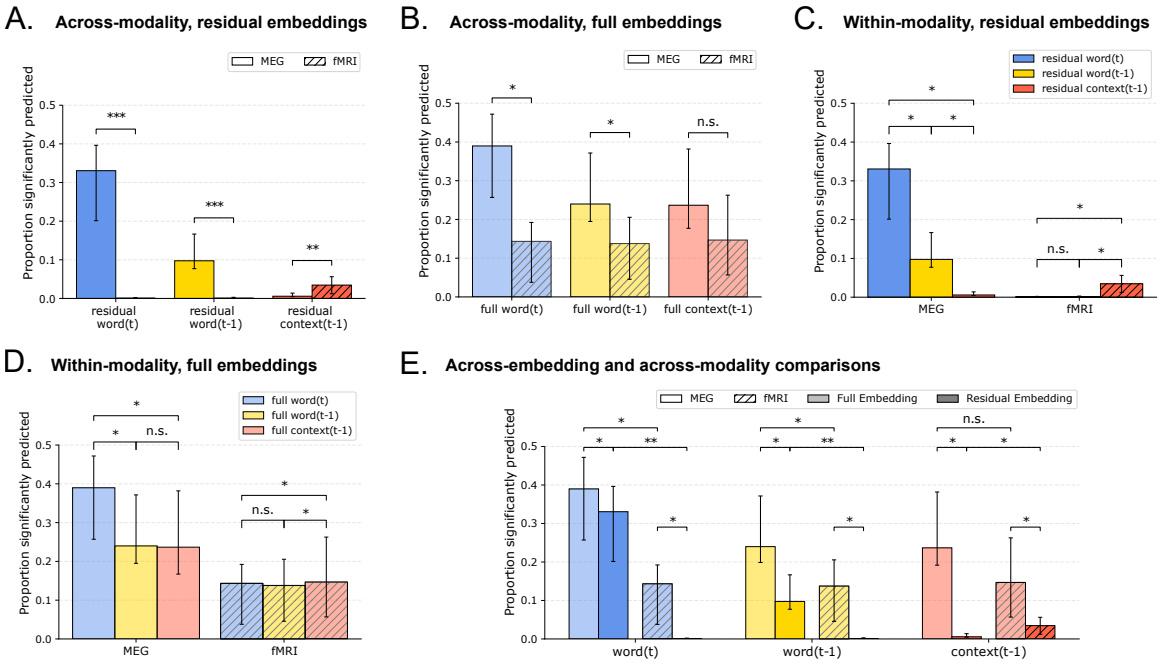


Figure 4: Direct comparisons of prediction performance of different meaning embeddings. Displayed are the median proportions across participants and the medians' 95% confidence intervals. Differences between modalities are tested for significance using a Wilcoxon rank-sums test. Differences within modality are tested using a Wilcoxon signed-rank test. All p-values are adjusted for multiple comparisons with the Holm-Bonferroni procedure at $\alpha = 0.05$. **(A)** Residual previous word, context, and current word embeddings predict fMRI and MEG with significant differences. **(B)** Full context embeddings do not predict fMRI and MEG with significant differences, while the full current word and previous word embeddings predict MEG significantly better than fMRI. **(C)** MEG and fMRI display a contrasting pattern of prediction by the residual embeddings. The current word residual best predicts MEG activity, significantly better than the previous word residual, which in turns predicts MEG significantly more than the context residual. In contrast, the context residuals significantly predict fMRI activity better than the previous and current word residuals. **(D)** Full previous word and context embeddings do not predict MEG significantly differently. **(E)** All full embeddings predict both fMRI and MEG significantly better than the corresponding residual embeddings.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

word has many possible explanations. One possible explanation is that a word might take longer to process and integrate into the composed meaning than the duration it is on the screen. Another possible explanation is that a word may constrain the processing of the word that follows it, highlighting its relevant properties and aiding with composition. The hypothesis that MEG recordings reflect the process of composition aligns well with a vast number of previous findings characterizing transient responses evoked by a stimulus that is difficult to integrate with the preceding context (Kutas and Federmeier, 2011; Kuperberg *et al.*, 2003; Kuperberg, 2007) and results showing that MEG recordings are better fit by a model constrained by the meaning of the immediately preceding words (Lyu *et al.*, 2019). Indeed, our results are not in disagreement with this literature – they do not show that MEG activity does not reveal word integration processes that depend on previous context. Instead, our results suggest that the representation of that previous context is not visible in MEG.

The observed difference in predicting fMRI and MEG recordings raises the hypothesis that the process of maintenance of the composed meaning does not rely on neural mechanisms that are thought to generate the MEG signal (such as synchronized current flow in pyramidal cell dendrites (Hall *et al.*, 2014)) but on some other mechanisms that are not visible in the MEG signal or might be indistinguishable from noise (e.g. unsynchronized neural firing), but that have enough metabolic demands to generate a BOLD response. Alternate possible explanations for the lack of predictability of MEG by supra-word meaning are that the representation of supra-word meaning may be too distributed to be captured by MEG due to its poor spatial resolution. However, we observe that the supra-word meaning predicts about $5 - 10\%$ of all cortical fMRI voxels across participants, mostly centered in the ATL and PTL. Thus, it is unlikely that no MEG sensor-timepoint is sensitive to this signal if it is detectable in the magnetic field changes. Further, MEG is known to be sensitive to neural activity that originates in the sulci, and since we find that the voxels that are sensitive to supra-word meaning are in the sulci, this explanation is even less likely. These results call for a new understanding of previous work that aims to study composition of sentence-level meaning using MEG as well as possibly other types of imaging modalities that rely on synchronized firing, such as EEG and ECoG. Our results suggest that observed increases in activity measured by these modalities during sentence reading (Fedorenko *et al.*, 2016; Hultén *et al.*, 2019) and improved fit by a model constrained by very recent context (Lyu *et al.*, 2019) are due to instantaneous integration processes rather than sentence-level meaning. Future work is needed to understand whether and how these imaging modalities can be used to study sentence-level meaning.

Our analysis depends on the degree to which the computational neural network we have chosen is able to represent composed meaning. Based on ELMo's competitive performance on downstream tasks (Peters *et al.*, 2018) and ability to capture complex linguistic structure (Tenney *et al.*, 2019), we believe that ELMo is able to extract some aspects of composed meaning. The degree to which this composed meaning reflects the one in the brain is an important question that we have only begun to study and needs further investigation. Secondly, our residual approach accounts only for the linear dependence between individual word embeddings and context embeddings. By construction, the internal state of the LSTM in ELMo contains non-linear dependencies on the input word vector and the previous LSTM state. It is possible however that some dimensions of the internal state of the ELMo LSTM corresponds to non-linear operations on the dimensions of the input vector alone, without

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a contribution from the previous internal state of the LSTM (see Materials and Methods for the LSTM equations). This non-linear transformation of the input word might not be removed by our residual procedure, and whether it aligns with processing of individual words in the brain is a question for future research.

The surprising finding that supra-word meaning is difficult to capture using MEG has implications for future neuroimaging research and applications where natural language is decoded from the brain. While high temporal imaging resolution is key to reaching a mechanistic level of understanding of language processing, our findings suggest that a modality other than MEG may be necessary to detect long-range contextual information. Further, the fact that an aspect of meaning can be predictive in one imaging modality and invisible in the other calls for caution while interpreting findings about the brain from one modality alone, as some parts of the puzzle are systematically hidden. Our results also suggest that the imaging modality may impact the ability to decode the contextualized meaning of words, which is central to brain-computer interfaces (BCI) that aim to decode attempted speech. Recent success in decoding speech from ECoG recordings (Makin *et al.*, 2020) is promising, but needs to be evaluated carefully with more diverse and naturalistic stimuli. Using BCI to decode speech in real life is complicated by the inherent uncertainty in decoding each word and the fact that the space of all possible utterances is not constrained. It is yet to be determined if word-level information conveyed by electrophysiology will be enough to decode a person's intent, or if the lack of supra-word meaning should be compensated in other ways.

Materials and Methods

fMRI data and preprocessing

We use fMRI data of 9 participants reading chapter 9 of *Harry Potter and the Sorcerer's Stone* (Rowling, 2012), collected and made available online by Wehbe *et al.* (2014b). Words were presented one at a time at a rate of 0.5s each. fMRI data was acquired at a rate of 2s per image, i.e. the repetition time (TR) is 2s. The images were comprised of $3 \times 3 \times 3mm$ voxels. The data for each participant was slice-time and motion corrected using SPM8 (Kay *et al.*, 2008), then detrended and smoothed with a 3mm full-width-half-max kernel. The brain surface of each participant was reconstructed using Freesurfer (Fischl, 2012), and a grey matter mask was obtained. The Pycortex software (Gao *et al.*, 2015) was used to handle and plot the data. For each participant, $25000 - 31000$ cortical voxels were kept.

MEG data and preprocessing

The same paradigm was recorded for 8 participants using MEG by the authors of Wehbe *et al.* (2014a) and shared upon our request. This data was recorded at 306 sensors organized in 102 locations around the head. MEG records the change in magnetic field due to neuronal activity and the data we used was sampled at 1kHz, then preprocessed using the Signal Space Separation method (SSS) (Taulu *et al.*, 2004) and its temporal extension (tSSS) (Taulu and Simola, 2006). The signal in every sensor was downsampled into 25ms non-overlapping time bins. For each of the 5176 word in the chapter, we therefore obtained a recording for 306

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sensors at 20 time points after word onset (since each word was presented for 500ms).

ELMo details

At each layer, for each word ELMo combines the internal representations of two independent LSTMs – a forward LSTM (containing information from previous words) and a backward LSTM (containing information from future words). We extracted context embeddings only from the forward LSTM in order to more closely match the participants, who have not seen the future words. For a word token t, the forward LSTM generates the hidden representation h_t^l in layer l using the following update equations:

$$\begin{aligned}\tilde{c} &= \tanh(w_c[h_{t-1}^l; h_t^{l-1}] + b_c),\\ c_t &= f_t \times c_{t-1} + i_t \times \tilde{c}_t,\\ h_t^l &= o_t \times \tanh(c_t),\end{aligned}$$

where b_c and w_c represent the learned bias and weight, and f_t, o_t, and i_t represent the forget, output, and input gates. The states of the gates are computed according to the following equations:

$$\begin{aligned}f_t &= \sigma(w_f[h_{t-1}^l; h_t^{l-1}] + b_f),\\ i_t &= \sigma(w_i[h_{t-1}^l; h_t^{l-1}] + b_i),\\ o_t &= \sigma(w_o[h_{t-1}^l; h_t^{l-1}] + b_o),\end{aligned}$$

where $\sigma(x)$ represents the sigmoid function and b_x and w_x represent the learned bias and weight of the corresponding gate. The learned parameters are trained to predict the identity of a word given a series of preceding words, in a large text corpus. We use a pretrained version of ELMo with 2 hidden LSTM layers provided by Gardner *et al.* (2018).

Obtaining full stimulus representations

We obtain a full ELMo word embedding (as opposed to a residual word embedding) for word w_n by passing word w_n through the pretrained ELMo model and obtaining the token-level embeddings (i.e. from layer 0) for w_n. If word w_n contains multiple tokens, we average the corresponding token-level embeddings and use this average as the final full word embedding. We obtain a full ELMo context embedding for word w_n by passing the most recent 25 words $(w_{n-24}, .., w_n)$ through the pretrained ELMo model and obtaining the embeddings from the first hidden layer (i.e. from layer 1) of the forward LSTM for w_n. If word w_n contains multiple tokens, we average the corresponding layer 1 embeddings and use this mean as the final full context embedding for word w_n. We use 25 words to extract the context embedding because it has been previously shown that ELMo and other LSTMs appear to reduce the amount of information they maintain beyond $20 - 25$ words in the past (Khandelwal *et al.*, 2018; Toneva and Wehbe, 2019).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Obtaining residual stimulus representations

We obtain three types of residual embeddings for each word at position t in the stimulus set: 1) residual context(t-1) embedding, 2) residual word(t-1) embedding, and 3) residual word(t) embedding. We compute all three types using the same general approach of training a regularized linear regression, but with inputs x_t and outputs y_t that change depending on the type of residual embedding. The steps to the general approach are the following, given an input x_t and output y_t:
Step 1: Learn a linear function g that predicts each dimension of y_t as a linear combination of x_t. We follow the same steps outlined in the training of function f in the encoding model. Namely, we model g as a linear function, regularized by the ridge penalty. The model is trained via four-fold cross-validation and the regularization parameter is chosen via nested cross-validation.
Step 2: Obtain the residual $y_t^{\prime} \triangleq y_t - \hat{g}(x_t)$, using the estimate of the g function learned above. This is the final residual stimulus representation.

For the residual context(t-1) embedding, the input x_t is the concatenation of the full word embeddings for the 25 consecutive words $w_{t-24}, ..., w_t$ and the output y_t is the full context(t-1) embedding. For the residual word(t-1) embeddings, the input x_t is the concatenation of the full context(t-1) embedding and the full word embeddings for the 24 consecutive words $w_{t-24}, ..., w_t$ that exclude the full word embedding for word(t-1) and the output y_t is the full word(t-1) embedding. For the residual word(t) embeddings, the input x_t is the the concatenation of the full context(t-1) embedding and the full word embeddings for the 24 consecutive words $w_{t-24}, ..., w_{t-1}$ and the output y_t is the full word(t) embedding.

Encoding model evaluation

We evaluate the predictions of each encoding model by computing the Pearson correlation between the held-out brain recordings and the corresponding predictions in the four-fold cross-validation setting. We compute one correlation value for each of the 4 cross-validation folds and report the average value as the final encoding model performance.

General encoding model training

For each type of embedding e_t, we estimate an encoding model that takes e_t as input and predicts the brain recording associated with reading the same words that were used to derive e_t. We estimate a function f, such that $f(e_t) = b$, where b is the brain activity recorded with either MEG or fMRI. We follow previous work (Sudre *et al.*, 2012; Wehbe *et al.*, 2014b,a; Nishimoto *et al.*, 2011; Huth *et al.*, 2016) and model f as a linear function, regularized by the ridge penalty. The model is trained via four-fold cross-validation and the regularization parameter is chosen via nested cross-validation.

fMRI Encoding Models

Ridge regularization is used to estimate the parameters of a linear model that predicts the brain activity y^i in every fMRI voxel i as a linear combination of a particular NLP embedding x. For each output dimension (voxel), the Ridge regularization parameter is

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

chosen independently by nested cross-validation. We use Ridge regression because of its computational efficiency and because of the results of Wehbe *et al.* (2015) showing that for fMRI data, as long as proper regularization is used and the regularization parameter is chosen by cross-validation for each voxel independently, different regularization techniques lead to similar results. Indeed, Ridge regression is indeed a common regularization technique used for building predictive fMRI (Mitchell *et al.*, 2008; Nishimoto *et al.*, 2011; Wehbe *et al.*, 2014b; Huth *et al.*, 2016).

For every voxel i, a model is fit to predict the signals $y^i = [y^i_1, y^i_2, \ldots, y^i_n]$, where n is the number of time points, as a function of the NLP embedding. The words presented to the participants are first grouped by the TR interval in which they were presented. Then, the NLP embedding of the words in every group are averaged to form a sequence of features $x = [x_1, x_2, \ldots, x_n]$ which are aligned with the brain signals. The models are trained to predict the signal at time t, y_t, using the concatenated vector z_t formed of $[x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}]$. The features of the words presented in the previous volumes are included in order to account for the lag in the hemodynamic response that fMRI records. Indeed, the response measured by fMRI is an indirect consequence of brain activity that peaks about 6 seconds after stimulus onset, and the solution of expressing brain activity as a function of the features of the preceding time points is a common solution for building predictive models (Nishimoto *et al.*, 2011; Wehbe *et al.*, 2014b; Huth *et al.*, 2016).

For each given participant and each NLP embedding, we perform a cross-validation procedure to estimate how predictive that NLP embedding is of brain activity in each voxel i. For each fold:

- The fMRI data Y and feature matrix $Z = z_1, z_2, \ldots z_n$ are split into corresponding train and validation matrices. These matrices are individually normalized (mean of 0 and standard deviation of 1 for each voxel across time), ending with train matrices Y^R and Z^R and validation matrices Y^V and Z^V.

- Using the train fold, a model w^i is estimated as:

$$\arg\min_{w^i} ||y^{R,i} - Z^R w^i|^2_2 + \lambda^i ||w^i||^2_2.$$

  A ten-fold nested cross-validation procedure is first used to identify the best λ^i for every voxel i that minimizes nested cross-validation error. w^i is then estimated using λ^i on the entire training fold.

- The predictions for each voxel on the validation fold are obtained as $p = Z^V w^i$.

The above steps are repeated for each of the four cross-validation folds and average correlation is obtained for each voxel i, NLP embedding, and participant.

MEG encoding models

MEG data is sampled faster than the rate of word presentation, so for each word, we have 20 times points recorded at 306 sensors. Ridge regularization is similarly used to estimate the parameters of a linear model that predicts the brain activity $y^{i,\tau}$ in every MEG sensor i

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

at time τ after word onset. For each output dimension (sensor/time tuple i, τ), the Ridge regularization parameter is chosen independently by nested cross-validation.

For every tuple i, τ, a model is fit to predict the signals $y^{i,\tau} = [y_1^{i,\tau}, y_2^{i,\tau}, \ldots, y_n^{i,\tau}]$, where n is the number of words in the story, as a function of NLP embeddings. We use as input the word vector x without the delays we used in fMRI because the MEG recordings capture instantaneous consequences of brain activity (change in the magnetic field). The models are trained to predict the signal at word t, $y_t^{i,\tau}$, using the vector x_t.

For each participant and NLP embedding, we perform a cross-validation procedure to estimate how predictive that NLP embedding is of brain activity in each sensor-timepoint i. For each fold:

- The MEG data Y and feature matrix $X = x_1, x_2, \ldots x_n$ are split into corresponding train and validation matrices and these matrices are individually normalized (to get a mean of 0 and standard deviation of 1 for each voxel across time), ending with train matrices Y^R and X^R and validation matrices Y^V and Z^V.
- Using the train fold, a model $w^{(i,\tau)}$ is estimated as:

$$\arg\min_{w^{(i,\tau)}} ||y^{(i,\tau),R} - X^R w^{(i,\tau)}|_2^2 + \lambda^{(i,\tau)} ||w^{(i,\tau)}||_2^2.$$

  A ten-fold nested cross-validation procedure is first used to identify the best $\lambda^{(i,\tau)}$ for every sensor, time-point tuple (i, τ) that minimizes the nested cross-validation error. $w^{(i,\tau)\ell}$ is then estimated using $\lambda^{(i,\tau)}$ on the entire training fold.
- The predictions for each sensor, time-point tuple (i, τ) on the validation fold are obtained as $p = X^V w^{(i,\tau)}$.

The above steps are repeated for each of the four cross-validation folds and an average correlation is obtained for each sensor location, time-point tuple (s, τ), each NLP embedding, and each participant.

Spatial Generalization Matrices

We introduce the concept of spatial generalization matrices, which tests whether an encoding model trained to predict a particular voxel can generalize to predicting other voxels. This approach can be applied to voxels within the same participant or in other participants. The purpose of this method is to test whether two voxels relate to specific representation of the input (e.g. NLP embedding) in a similar way. If an encoding model for a particular voxel is able to significantly predict a different voxel's activity, we conclude that the two voxels process similar information with respect to the input of the encoding model.

For each pair of voxels (i, j), we first follow our general approach of training an encoding model to predict voxel i as a function of a specific stimulus representation, described above, and test how well the predictions of the encoding model correlate with the activity of voxel j. We do this for all pairs of voxels in the PTL and ATL across all 9 participants. We finally normalize the resulting performance at predicting voxel j by dividing it by the performance at predicting test data from voxel i, that was heldout during the training process. The significance of the performance of the encoding model on voxel j is evaluated using a permutation test, described below.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Permutation tests

Significance of the degree to which a single voxel or a sensor-timepoint is predicted is evaluated based on a standard permutation test. To conduct the permutation test, we block-permute the predictions of a specific encoding model within each of the four cross-validation runs and compute the correlation between the block-permuted predictions and the corresponding true values of the voxel/sensor-timepoint. We use blocks of 5TRs in fMRI (corresponding to 20 presented words) and 20 words in MEG in order to retain some of the auto-regressive structure in the permuted brain recordings. We conduct 1000 permutations and calculate the number of times the resulting mean correlation across the four cross-validation folds of the permuted predictions is higher than the mean correlation from the original unpermuted predictions. The resulting p-values for all voxels/sensor-timepoints/time-windows are FDR corrected for multiple comparisons using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Chance proportions of ROI/timewindows predicted significantly

To establish whether a significant proportion of an ROI/timewindow is predicted by a specific encoding model, we contrast the proportion of the ROI/timewindow that is significantly explained by the encoding model with a proportion of the ROI/timewindow that is significantly explained by chance. We do this for all proportions of the same ROI/timewindow across participants, using a Wilcoxon signed-rank test. We compute the proportion of an ROI/timewindow that is significantly predicted by chance using the permutation tests described above. For each permutation k, we compute the p-value of each voxel in this permutation according to its performance with respect to the other permutations. Next for each ROI/timewindow, we compute the proportion of this ROI/timewindow with p-values$<$ 0.01 after FDR correction, for each permutation. The final chance proportion of an ROI/time-window for a specific encoding model and participant is the average chance proportion across permutations.

Confidence intervals

We use an open-source package (Sheppard *et al.*, 2020) to compute the 95% bias-corrected confidence intervals of the median proportions across participants. We use bias-corrected confidence intervals (Efron and Tibshirani, 1994) to account for any possible bias in the sample median due to a small sample size or skewed distribution (Miller, 1988).

Experiments revealing shared information among NLP embeddings

For each of the 3 NLP embedding types (i.e. context(t-1) embedding, word(t-1) embedding, word(t) embedding), we train an encoding model taking as input each NLP embedding and predicting as output the word embedding for word(i), where $i \in [t-6, t+2]$. We evaluate the predictions of the encoding models using Pearson correlation, and obtain an average correlation over the four cross-validation folds.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments

The authors thank Erika Laing and Daniel Howarth for help with data collection and preprocessing, and Michael J. Tarr for helpful feedback on the manuscript. This research was supported in part by start-up funds in the Machine Learning Department at Carnegie Mellon University, the Google Faculty Research Award and the Air Force Office of Scientific Research through research grants FA95501710218 and FA95502010118.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, **57**(1), 289–300.

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, **19**(12), 2767–2796.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Fedorenko, E. and Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in cognitive sciences*, **18**(3), 120–126.

Fedorenko, E., Hsieh, P.-J., Nieto-Castanon, A., Whitfield-Gabrieli, S., and Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, **104**(2), 1177–1194.

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., and Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, **113**(41), E6256–E6262.

Fischl, B. (2012). Freesurfer. *Neuroimage*, **62**(2), 774–781.

Frankland, S. M. and Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, **112**(37), 11732–11737.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, **91**(4), 1357–1392.

Fyshe, A., Sudre, G., Wehbe, L., Rafidi, N., and Mitchell, T. M. (2019). The lexical semantics of adjective–noun phrases in the human brain. *Human brain mapping*, **40**(15), 4457–4469.

Gao, J. S., Huth, A. G., Lescroart, M. D., and Gallant, J. L. (2015). Pycortex: an interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, **9**, 23.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hagoort, P. (2020). The meaning-making mechanism (s) behind the eyes and between the ears. *Philosophical Transactions of the Royal Society B*, **375**(1791), 20190301.

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., and Dale, A. M. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *Neuroimage*, **17**(3), 1101–1116.

Hall, E. L., Robson, S. E., Morris, P. G., and Brookes, M. J. (2014). The relationship between meg and fmri. *Neuroimage*, **102**, 80–91.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Hultén, A., Schoffelen, J.-M., Uddén, J., Lam, N. H., and Hagoort, P. (2019). How the brain makes sense beyond the processing of single words–an meg study. *Neuroimage*, **186**, 586–594.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., Gallant, J. L., Heer, W. a. D., Griffiths, T. L., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, **532**(7600), 453–458.

Jain, S. and Huth, A. (2018). Incorporating context into language encoding models for fmri. In *Advances in neural information processing systems*, pages 6628–6637.

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, **452**(7185), 352.

Khandelwal, U., He, H., Qi, P., and Jurafsky, D. (2018). Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294.

King, J.-R. and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, **18**(4), 203–210.

Kuperberg, G. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, **1146**, 23–49.

Kuperberg, G. R., Holcomb, P. J., Sitnikova, T., Greve, D., Dale, A. M., and Caplan, D. (2003). Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *Journal of Cognitive Neuroscience*, **15**(2), 272–293.

Kutas, M. and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, **62**, 621–647.

Lyu, B., Choi, H. S., Marslen-Wilson, W. D., Clarke, A., Randall, B., and Tyler, L. K. (2019). Neural dynamics of semantic composition. *Proceedings of the National Academy of Sciences*, **116**(42), 21318–21327.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Makin, J. G., Moses, D. A., and Chang, E. F. (2020). Machine translation of cortical activity to text with an encoder–decoder framework. Technical report, Nature Publishing Group.

Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, **14**(3), 539.

Mitchell, T., Shinkareva, S., Carlson, A., Chang, K., Malave, V., Mason, R., and Just, M. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, **320**(5880), 1191–1195.

Nishimoto, S., Vu, A., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*.

Pallier, C., Devauchelle, A., and Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, **108**(6), 2522–2527.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Pylkkänen, L. (2020). Neural basis of basic composition: what we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B*, **375**(1791), 20190299.

Rowling, J. (2012). *Harry Potter and the Sorcerer's Stone*. Harry Potter US. Pottermore Limited.

Sheppard, K., Khrapov, S., Lipták, G., mikedeltalima, Capellini, R., Hugle, esvhd, Fortin, A., JPN, Adams, A., jbrockmendel, Rabba, M., Rose, M. E., Rochette, T., RENE-CORAIL, X., and syncoding (2020). bashtage/arch: Release 4.15.

Skeide, M. A. and Friederici, A. D. (2016). The ontogeny of the cortical language network. *Nature Reviews Neuroscience*, **17**(5), 323–332.

Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., and Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, **62**, 451–463.

Taulu, S. and Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in medicine and biology*, **51**(7), 1759.

Taulu, S., Kajola, M., and Simola, J. (2004). Suppression of interference and artifacts by the signal space separation method. *Brain topography*, **16**(4), 269–275.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S., Das, D., *et al.* (2019). What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Toneva, M. and Wehbe, L. (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.

Visser, M., Jefferies, E., and Lambon Ralph, M. (2010). Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *Journal of cognitive neuroscience*, **22**(6), 1083–1094.

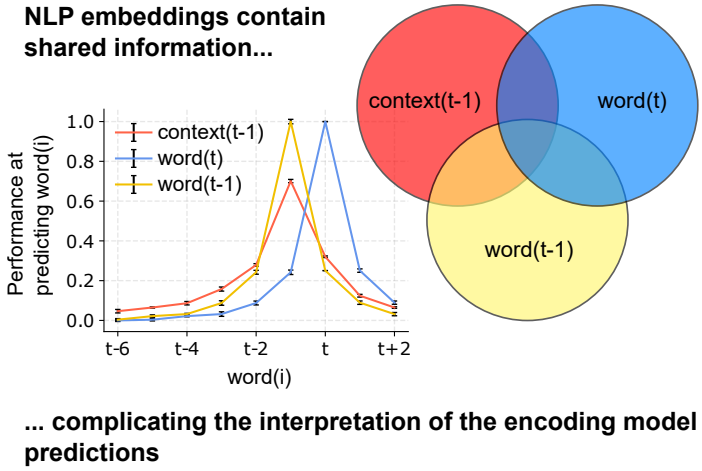
Wehbe, L., Vaswani, A., Knight, K., and Mitchell, T. (2014a). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., and Mitchell, T. (2014b). Simultaneously uncovering the patterns of brain regions involved in different story reading Subprocesses. *PloS one*, **9**(11), e112575.

Wehbe, L., Ramdas, A., Steorts, R. C., and Shalizi, C. R. (2015). Regularized brain reading with shrinkage and smoothing. *Annals of Applied Statistics*, **9**(4), 1997–2022.

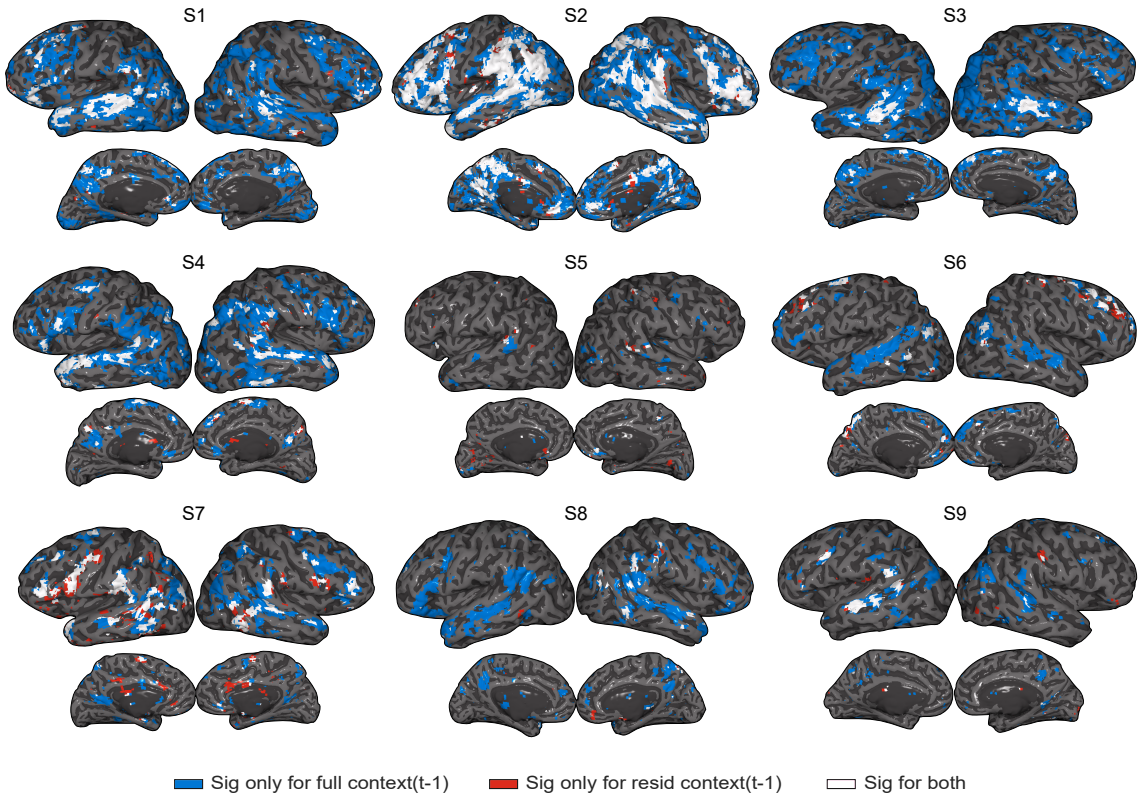
bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary



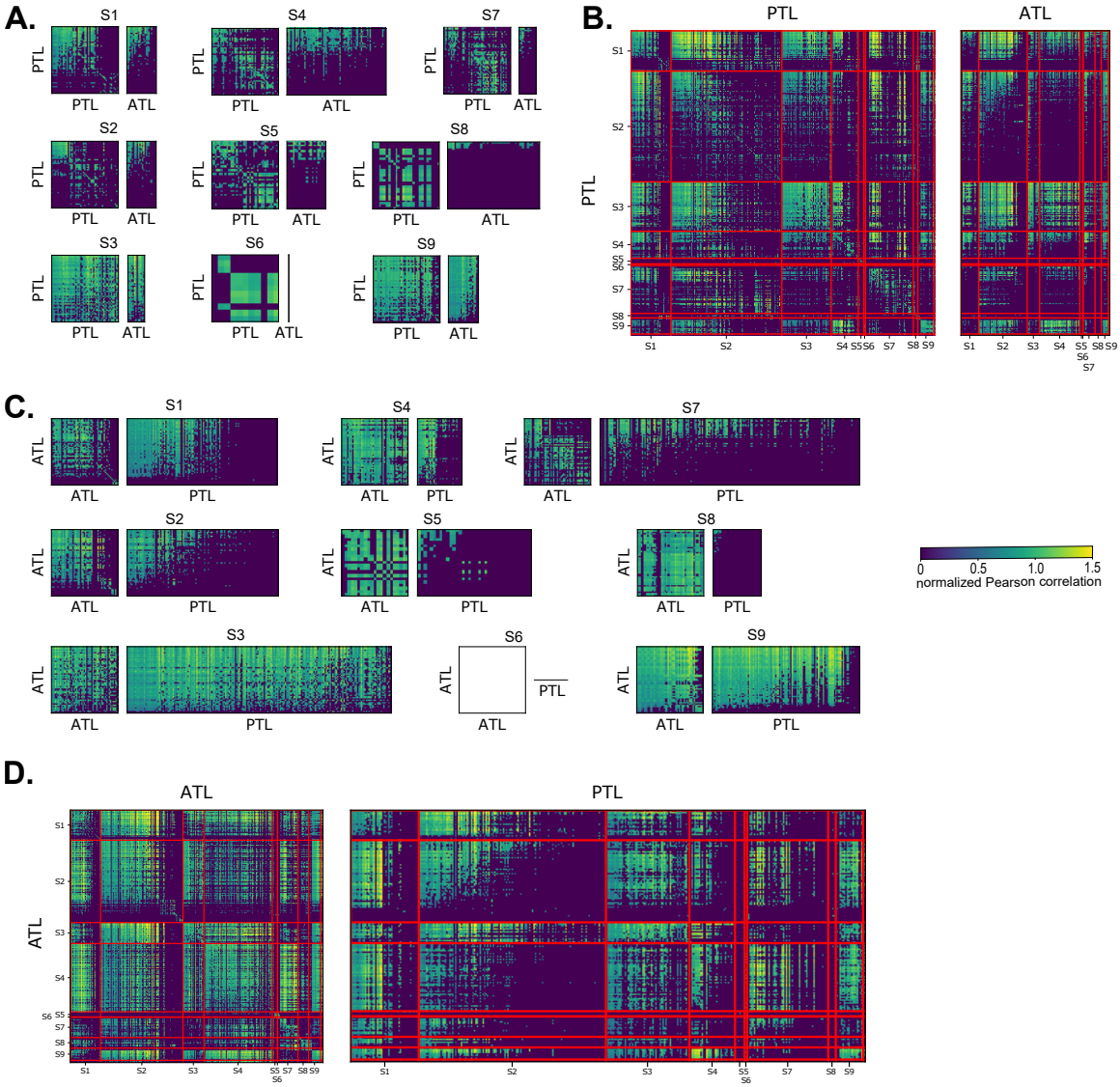
Supplementary Figure 1: Context and word embeddings share information. The performance of the context and word embeddings at predicting the words at surrounding positions is plotted for different positions. The context embedding has information about up to 6 past words, and word embeddings have information about embeddings of surrounding words. To isolate the information exclusively related to the meaning of an individual word, it is necessary to account for this shared information.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



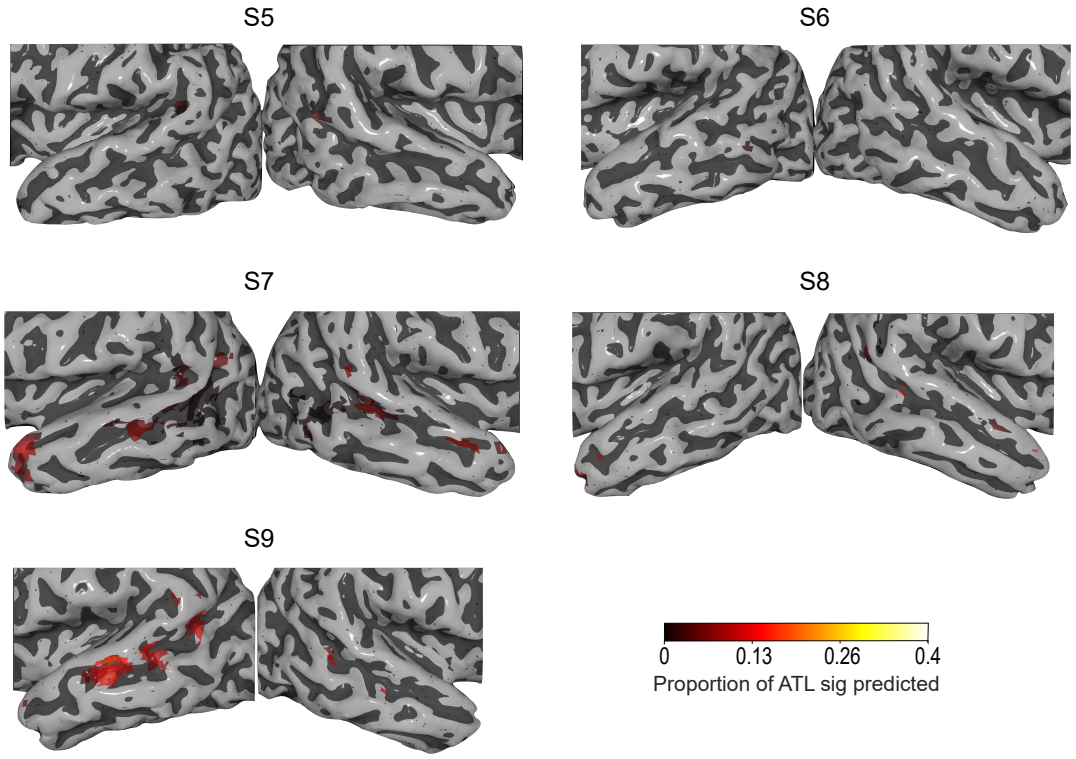
Supplementary Figure 2: Qualitative visualization of the voxels that are significantly predicted by the full contextualized representation (in blue), the residual contextualized representation (in red), or both (in white). A voxel is determined to be significantly predicted through a permutation test and FDR correction for multiple comparisons at the 0.01 level. Large parts of the language system, spanning the temporal cortex and the inferior frontal cortex, are significantly predicted by the full context embeddings. The voxels significantly predicted by the context residual are largely a subset of those predicted by the full context embeddings.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 3: Spatial Generalization Matrices for all 9 participants. Models trained to predict PTL voxels are used to predict PTL and ATL voxels (within-participant (**A**), and across-participants (**B**)). Models trained to predict ATL voxels are used to predict ATL and PTL voxels (within-participant (**C**), and across-participants (**D**)). Note that the block diagonal matrices of the across-participants correlations (in **B**/**D**) are equivalent to the plots in **A**/**C**. Only voxels that are significantly predicted by the context residual are included in this analysis. Note that the participant S6 does not have any significantly predicted voxels in the ATL. Correlations are normalized by dividing the performance of a model trained on voxel i at predicting the target voxel j by the performance of a model trained on the target voxel j. PTL cross-voxel correlations form two clusters: voxels in a cluster can predict each other but not the other cluster's voxels. Across participants, only one of these clusters has voxels that predict ATL voxels.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.28.316935; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 4: Performance of encoding models trained on ATL and PTL voxels at predicting other participants' ATL for the remaining 5 participants. All participants who have more than a few significantly predicted voxels (6 out of 9 participants) show a cluster of predictive voxels in the pSTS.