Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines

Ariel Goldstein[1,2☩], Zaid Zada[1*], Eliav Buchnik[2*], Mariano Schain[2*], Amy Price[1*], Bobbi Aubrey[1,3*], Samuel A. Nastase[1*], Amir Feder[2*], Dotan Emanuel[2*], Alon Cohen[2*], Aren Jansen[2*], Harshvardhan Gazula[1], Gina Choe[1,3], Aditi Rao[1,3], Catherine Kim[1,3], Colton Casto[1], Lora Fanda[3], Werner Doyle[3], Daniel Friedman[3], Patricia Dugan[3], Roi Reichart[5], Sasha Devore[3], Adeen Flinker[3,] Liat Hasenfratz[1], Avinatan Hassidim[2], Michael Brenner[2,4], Yossi Matias[2], Kenneth A. Norman[1], Orrin Devinsky[3], Uri Hasson[1,2]

[1]Department of Psychology and the Neuroscience Institute, Princeton University, Princeton, NJ

[2]Google Research

[3]New York University School of Medicine, New York, NY

[4]School of Engineering and Applied Science, Harvard University, Boston, MA

[5]Faculty of Industrial Engineering and Management, Technion, Israel Institute of Technology

* Equal contribution

☩Corresponding author: ariel.y.goldstein@gmail.com

Abstract

Departing from traditional linguistic models, advances in deep learning have resulted in a new type of predictive (autoregressive) deep language models (DLMs). These models are trained to generate appropriate linguistic responses in a given context using a self-supervised prediction task. We provide empirical evidence that the human brain and autoregressive DLMs share two computational principles: 1) both are engaged in continuous prediction; 2) both represent words as a function of the previous context. Behaviorally, we demonstrate a match between humans and DLM's next-word predictions given sufficient contextual windows during the processing of a real-life narrative. Neurally, we demonstrate that the brain, like autoregressive DLMs, constantly predicts upcoming words in natural speech, hundreds of milliseconds before they are perceived. Finally, we show that DLM's contextual embeddings capture the neural representation of context-specific word meaning better than arbitrary or static semantic embeddings. Our findings suggest that autoregressive DLMs provide a novel and biologically feasible computational framework for studying the neural basis of language.

Introduction

The outstanding success of deep language models (DLMs) is striking from a theoretical and practical perspective because they have emerged from a very different scientific paradigm than traditional psycholinguist models[1]. In traditional psycholinguistic approaches, human language is explained with interpretable models that combine symbolic elements (e.g., nouns, verbs, adjectives, adverbs) with rule-based operations [2,3]. In contrast, DLMs are trained (i.e., *learn*) to perform tasks from real-world textual examples "in the wild," with minimal or no prior knowledge about the structure of language. In this paper, we focus on two core features of autoregressive DLMs: (1) they rely on a simple, yet highly effective, self-supervised task, with the sole aim to predict the next word based on preceding words (i.e., *context)*; and (2) they encode the unique meaning of each word based on the preceding context [4–8]. Autoregressive DLMs do not parse words into parts of speech or apply explicit syntactic transformations but rather learn to encode a sequence of words into a numerical vector, termed a contextual embedding, from which the model decodes the next word. After learning, the next-word prediction principle allows the generation of well-formed texts in entirely new contexts never seen during training [9,10,1].

While autoregressive DLMs and the contextual embeddings they learn have proven to be extremely effective in capturing the structure of language [11,5,12], it is unclear if the mechanisms underlying language processing in autoregressive DLMs are related to the way the human brain processes language. Past research has leveraged language models and machine learning to extract information about semantic representation structure in the brain [13–19]. But such studies did not view DLMs as feasible cognitive models for how the human brain is coding language. Some theoretical and empirical papers, however, have recently begun to search for shared computational principles between DLMs and the brain's representation of language [20,21]. The current paper tests whether two core computational principles central to autoregressive DLMs—next-word prediction and context-specific representation of meaning—are deployed by the brain as it processes natural language.

The claim that the brain is constantly engaged in predicting the incoming input is fundamental to numerous predictive coding theories [22–26]. However, even after decades of research, behavioral and neural evidence for the brain's propensity to predict upcoming words in natural language has remained indirect. On the behavioral level, the ability to predict upcoming words has been mostly tested with highly-controlled sentence stimuli (i.e., the cloze procedure [27–31]). Thus, we still do not know how accurate listeners actually predict words in open-ended natural contexts. The evidence of word prediction on the neural level is largely based on signals associated with prediction error detected 300 to 600 ms **after** word onset [32–36]. Post-word-onset prediction error signals, however reliable, cannot provide direct evidence of an active prediction of the upcoming words before they are perceived, as they can be attributed to any number of factors such as surprise, semantic relatedness, and attentional demands [37,38].

In the first section of the paper, we test whether the brain, like autoregressive DLMs, predicts upcoming words while listening to natural speech. We provide novel behavioral and neural

evidence for the spontaneous predictions of upcoming words **before** they are perceived, as the brain processes natural speech. Behaviorally, we found that, given a sufficient context window, an autoregressive DLM (GPT2) generates very similar next-word predictions to humans in a natural context. At the neural level, we leveraged high-precision electrocorticographic (ECoG) recordings to demonstrate that the brain spontaneously predicts the meaning of forthcoming words, even hundreds of milliseconds, before the words are perceived. Moreover, our analysis demonstrates that the neural signals before word onset are better modeled by the predicted words than the actually perceived ones.

In the second section of the paper, we test whether the brain, like autoregressive DLMs, encodes the unique, context-specific meaning of words based on the sequence of prior words. For that purpose, we extracted the contextual embeddings from an autoregressive DLM (GPT2) and used them to model the neural activity for each word in the story. Our results demonstrate that contextual embeddings improve our ability to model neural responses in multiple brain areas and at multiple time points to words in natural spoken language. Shuffling embeddings for occurrences of the same word abolishes this effect, demonstrating that the brain encodes the same word differently based on their context. Furthermore, the contextual embeddings can be used to predict each word's identity from cortical activity before word onset. Together, our findings provide compelling evidence for shared core computational principles, of prediction and contextual representation, between autoregressive DLMs and the human brain, and support a new modeling framework for studying the neural basis of the human language faculty.

Results

Section I: The brain spontaneously predicts words before they are perceived in natural language

Comparison of next-word prediction behavior in autoregressive DLMs and humans
We developed a novel sliding-window behavioral protocol to directly quantify humans' ability to predict every word in a natural context (Fig. 1A-B). 50 participants proceeded word-by-word through a 30-minute transcribed podcast ("Monkey in the Middle", *this American Life* podcast[39]) and provided a prediction of each upcoming word. The procedure yields 50 predictions for each of the story's 5113 words (see Fig. 1C, and Materials and Methods). We calculated a mean prediction performance for each word in the narrative, which we refer to as "predictability score" (Fig. 1D). A predictability score of 100% indicates that all participants correctly guessed the next word and a predictability score of 0% indicates that no participant predicted the upcoming word. This allows us to address the following questions: First, how good are humans at next-word prediction? Second, how much do human predictions align with DLM predictions?

Word-by-word behavioral prediction during a natural story
Participants were able to predict many upcoming words in a complex and unfamiliar story (mean predictability score = 28%, SE = 0.5%). The predictability score for blind guessing the most frequent word in the text ("the") was 6%. About 600 words had a predictability score higher than 70%. Interestingly, high predictability was not confined to the last words in a sentence and applied to words from all parts of speech (21.44% nouns, 14.64% verbs, 41.62% functions words, 4.35% adjectives, adverbs, and 17.94% other). This suggests that humans are proficient in predicting upcoming words in real-life contexts when asked to do so.

Comparing human and DLM next-word probabilities
Autoregressive DLMs learn how to generate well-formed linguistic outputs by improving their ability to predict the next word in natural linguistic contexts. We compared human and DLM ability to predict the same words in the podcast as a function of prior context. For each word in the transcript, we extracted the prediction probability assigned by an autoregressive DLM (GPT2) as a function of context (maximum context window of 1024 tokens). For example, GPT2 assigned a probability of 0.82 for the upcoming word "*monkeys*" when it received the preceding words in the story as contextual input: "So after two days of these near misses he changed strategies. He put his camera on a tripod and threw down some cookies to try to entice the _____." Human predictability scores and GPT2 estimations of predictability were highly correlated (Fig. 1E, $r = .79$, $p < .001$). This suggests that GPT2's and humans' next-word predictions are similar in natural contexts.

Prediction as a function of contextual window size
In natural comprehension (e.g., listening to or reading a story), predictions for upcoming words are influenced by information accumulated over multiple timescales: from the most recent words to the information gathered over multiple paragraphs [40]. We tested if GPT2's predictions would improve as a function of context window as it does in humans. To that end, we varied GPT2's input window size (from two tokens up to 1024 tokens) and examined how contextual window

size impacted the correlation with human behavior. The correlation between human and GPT2 word predictions improved as the contextual window increased (from $r = .46$, $p < .001$ at two-word context to an asymptote of $r = .79$ at 100-word context; Fig. 1F). This suggests that GPT2's predictions become more similar to human predictions as the contextual window size increases.

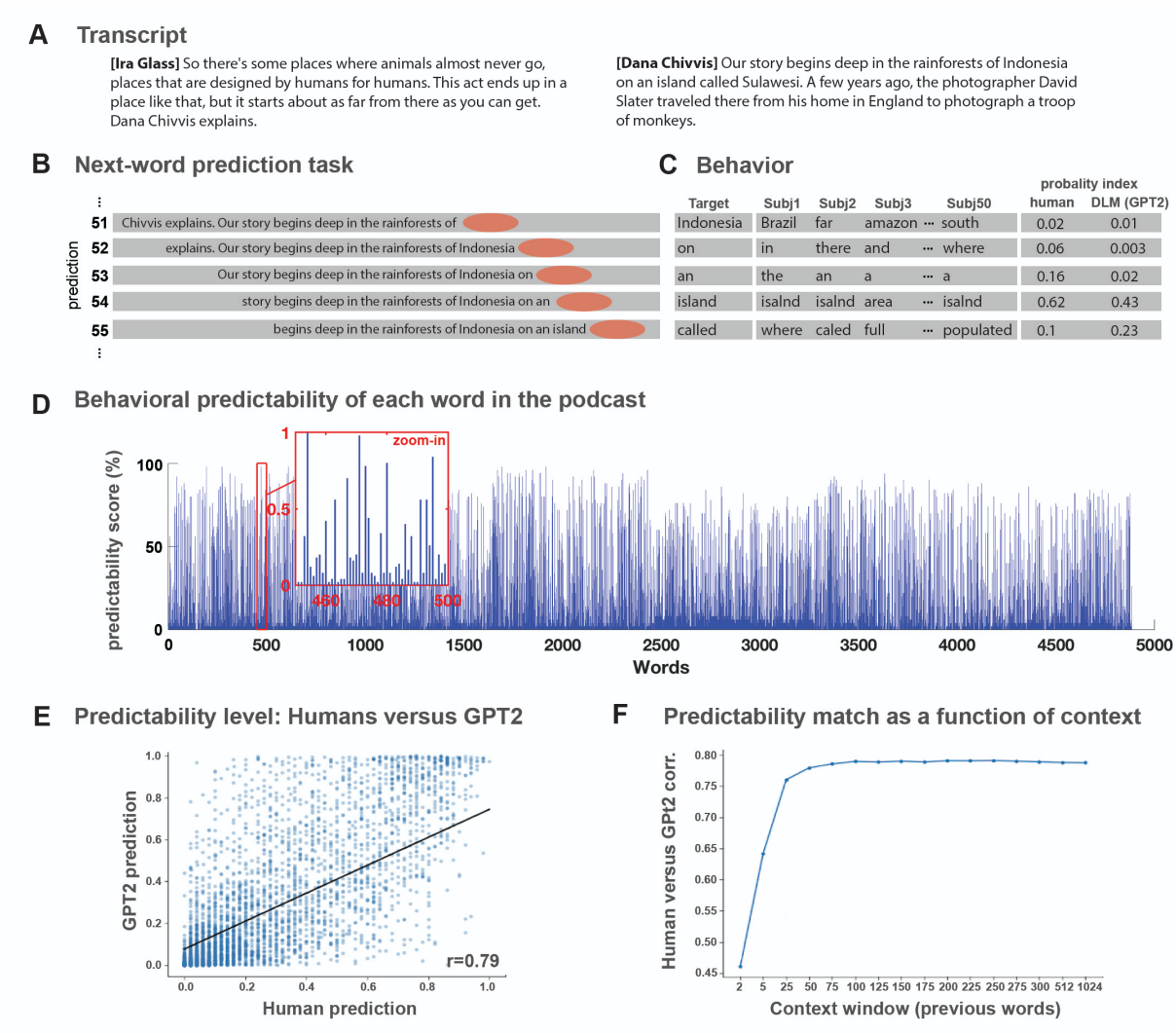

Figure 1. Behavioral assessment of the human ability to predict forthcoming words in a natural context. A) The stimulus was transcribed for the behavioral experiment. **B)** A 10-word sliding window was presented in each trial, and participants were asked to type their prediction of the next word. Once entered the correct word is presented, the window slides forward by one word. **C)** For each word, we calculated the proportion of participants that predicted the forthcoming word correctly. **D)** Human's predictability scores across words. **E)** Human's predictability scores versus GPT2's predictability scores for each upcoming word in the podcast. **F)** Correlation between human predictions and GPT2 predictions (as reported in panel D) for different context windows lengths ranging from 2–1024 preceding tokens.

Neural evidence for spontaneous next-word prediction

The behavioral study indicates that listeners can predict the upcoming next word in a natural open-ended context when explicitly instructed to do so in a stop-and-predict task. Furthermore, human predictions and autoregressive DLM predictions are matched in natural contexts. Next, we asked whether the human brain, like autoregressive DLM, is constantly engaged in spontaneous next-word prediction before word onset, even when it is not explicitly instructed to do so. To that end, we recorded electrocorticography (ECoG) signals from eight epileptic patients who volunteered to participate in the study (see Fig. 2A for a map of all electrodes). All participants listened to the same spoken story used in the behavioral experiment. In contrast to the behavioral study, however, participants engaged in free listening—with no explicit instructions to predict upcoming words. Comprehension was verified using a post-listening questionnaire.

Across participants, we had better coverage in the left hemisphere (917 electrodes) than in the right hemisphere (233 electrodes). Thus, we focus on language processing in the left hemisphere, but we also present results for the right hemisphere in supplementary materials for exhaustiveness. The raw signal is preprocessed to reflect high-frequency broadband (75–200Hz) power activity (for the full preprocessing procedure, please see Material and Methods).

We provide multiple pieces of evidence that the brain, like autoregressive DLMs, is spontaneously engaged in next-word prediction before word onset. To focus solely on the pre-onset prediction of the next word and not on the additional contextual information coded in contextual embedding (which we describe in the second section of this paper), we began with static semantic embeddings derived from the GloVe model[41]. Like GloVe and word2vec, static semantic embeddings produce a unique (50-dimensional) vector for every word in the lexicon[42]. For example, the word “monkey” receives a unique vector, which does not change across the different contexts it appears in throughout the podcast. This property of static embeddings allowed us to search for evidence for active prediction of each word before word onset irrespective of its context. Using static GloVe embeddings, we demonstrate that the brain contains information about upcoming words before they are perceived. Furthermore, we show that predictable words are encoded earlier than unpredictable words. Finally, we show that the neural activity before word onset contains information about the expected words, even when these predictions do not match the perceived words’ identity.

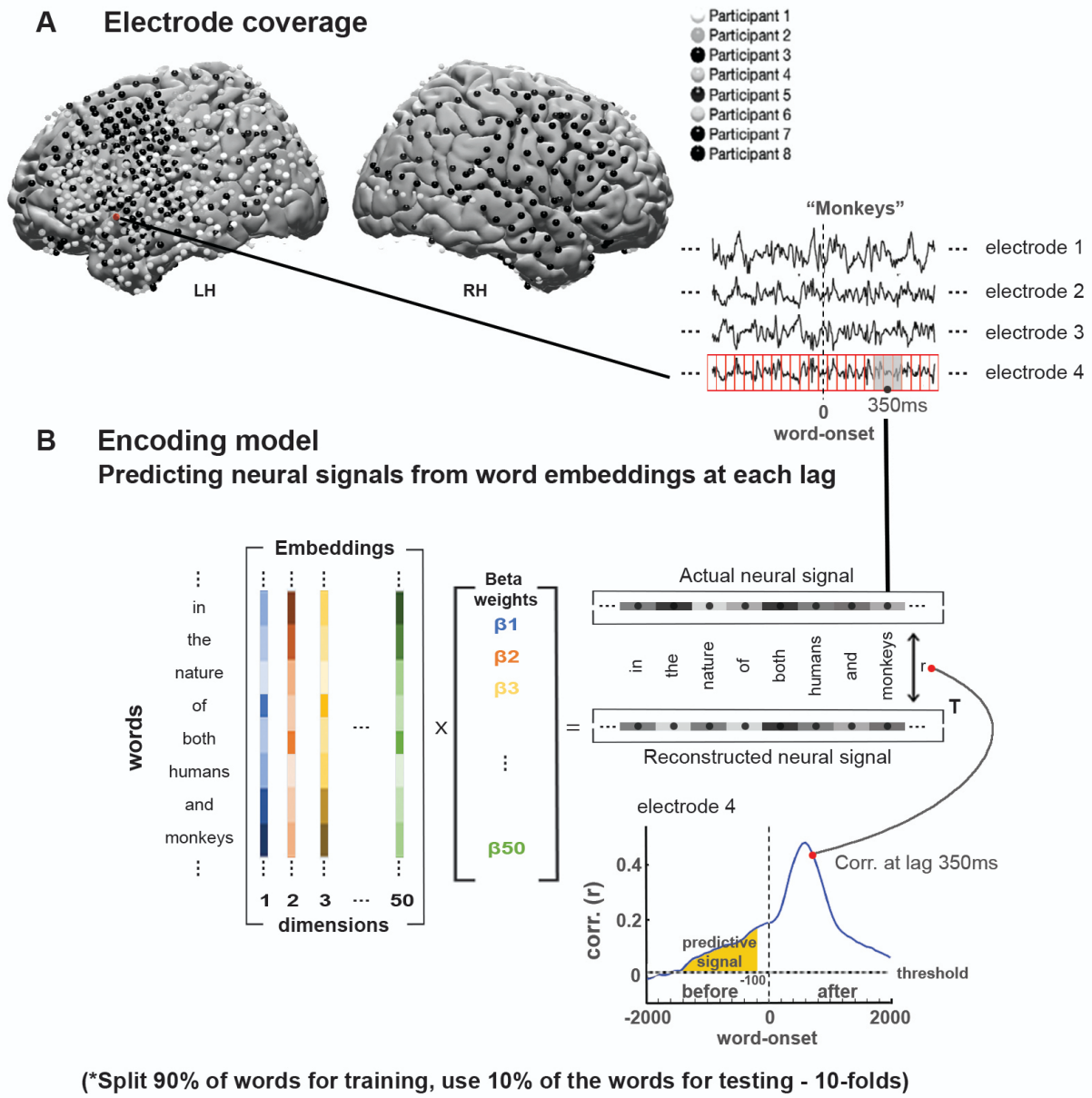

Figure 2. Linear encoding model used to predict the neural responses to each word in the narrative before and after word-onset. A) Brain coverage consisted of 1086 electrodes (eight participants). The words are aligned with the neural signal; each word's onset (moment of articulation) is at lag 0. Responses are averaged over a window of 200 ms and provided as input to the encoding model. **B)** A series of 50 coefficients corresponding to the features of the word embeddings is learned using linear regression to predict the neural signal across words from the assigned embeddings. The model was evaluated by computing the correlation between the reconstructed signal and the actual signal for the test words. This procedure was repeated for each lag and each electrode, using a 25 ms sliding window. The dashed horizontal line indicates the statistical threshold ($p < .01$ corrected for multiple comparisons). Lags of -100 ms or more preceding word onset contain only neural information sampled before the word was perceived (yellow color).

Localizing neural responses to natural speech using static embeddings
First, we used a linear encoding model and static semantic embeddings (GloVe) to localize electrodes containing reliable responses to single words in the narrative (see Fig. 2B and Materials and Methods). We use linear regression to learn a linear mapping between the GloVe embedding and the neural responses to each word (see Fig. 2B, and Materials and Methods section for details). GloVe embeddings capture some semantic and syntactic properties of language. To test if the embedded semantic relationships between words would improve the ability to model the neural activity to words, we also trained encoding models to predict neural responses using 50-dimensional static arbitrary embeddings, randomly sampled from a uniform [-1,1] distribution. This analysis provides us with a baseline measure capturing word identity deprived of any information about statistical relations among words.

The GloVe-based encoding model revealed consistent neural responses to words in a set of electrodes in auditory and language areas (Fig. 3A). The model identified 129 electrodes in the left hemisphere (LH) with significant correlations (after correction for multiple comparisons, see Materials and Methods for details). Electrodes were found in early auditory areas, motor cortex, and language areas (see Fig. 3A for LH electrodes, and supplementary Fig. S1 for right hemisphere (RH) electrodes; see Fig. S2 for single electrodes encoding models). GloVe embedding-based encoding models outperformed arbitrary embedding-based encoding models (Fig. 3B-C). Crucially, the improvement cannot be attributed to the GloVe vector space's wholesale geometrical properties, as it was abolished when we shuffled the assignment of the GloVe embeddings to words. thus removing the relational linguistic information from the model (Fig. S3). The improvement in modeling the neural responses using semantic embeddings is in line with recent results from other studies [13,43,44].

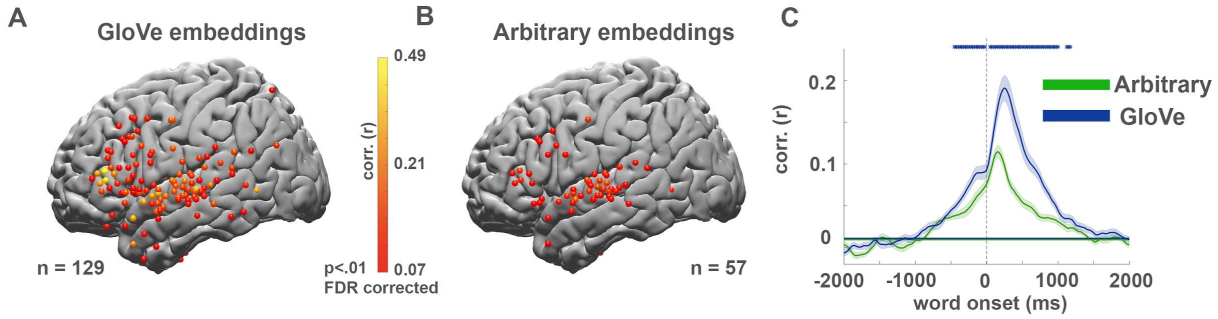

Figure 3. Static (GloVe) embeddings outperform arbitrary embeddings in predicting neural responses to words before and after word-onset. A) Electrodes with significant correlation at the peaked lag between predicted and actual word responses for semantic embeddings (GloVe). **B)** A weaker ability to predict neural responses to words for arbitrary embeddings. **C)** Average of the encoding models across the set of significant electrodes in 3B, separately for encoding based on arbitrary embedding model (green) and static semantic embedding model (blue). The standard error bands indicate the standard error of the encoding models across electrodes. The horizontal lines specify the statistical threshold. Blue asterisks indicate lags for which GloVe embeddings significantly outperform arbitrary embeddings ($p < .01$, nonparametric permutation test, FDR corrected).

Encoding neural responses before word onset
In the behavioral experiment (Fig. 1), we demonstrated people's capacity to predict upcoming words in the story. Next, we tested whether the neural signals also contain information about words before they are perceived (i.e., before word-onset).

The encoding model yielded significant correlations with predicted neural responses to upcoming words up to 800 ms before word-onset (Fig. 3C). The results were obtained for both semantic and arbitrary embeddings. Peak encoding performance for both arbitrary and semantic static embeddings was observed 150–200 ms after word onset (lag 0), but the models performed above chance up to 800 ms before word onset. This supports the claim that the brain deploys predictive semantic information about upcoming words' meaning before they are perceived. The significant encoding before word onset cannot be attributed to correlations between adjacent word embeddings in the story or to the existence of bigrams, as it was also apparent with arbitrary embeddings which do not contain information about the relationship among words. To further control the correlation among semantic static embeddings, we demonstrate that the significant encoding before word onset holds even after removing the previous GloVe embedding from the current GloVe embedding (Fig. S4). The encoding results using GloVe embeddings were replicated using 100-dimensional static embeddings from word2vec (Fig. S5). Together, these results indicate that the brain spontaneously (and accurately) predicts upcoming words before they are perceived.

Encoding of predictable and unpredictable words
Following the claim that the encoding results before word onset are related to next-word prediction, we expected predictable words to be encoded from the neural signals earlier than unpredictable words (Fig. 4A-B). To test this, we split the words according to humans and GPT2 predictability scores (see Fig. 1). The words with top-third predictability scores were classified as predictable, and bottom third scores were unpredictable. We trained the encoding models separately on the predictable and unpredictable groups using GloVe embeddings. The encoding model better predicted neural responses for predictable words than for unpredictable words at earlier time points relative to word-onset. Similar results were obtained when the set of predictable and unpredictable words were defined using human and GPT predictability scores (Fig. 4A-B).

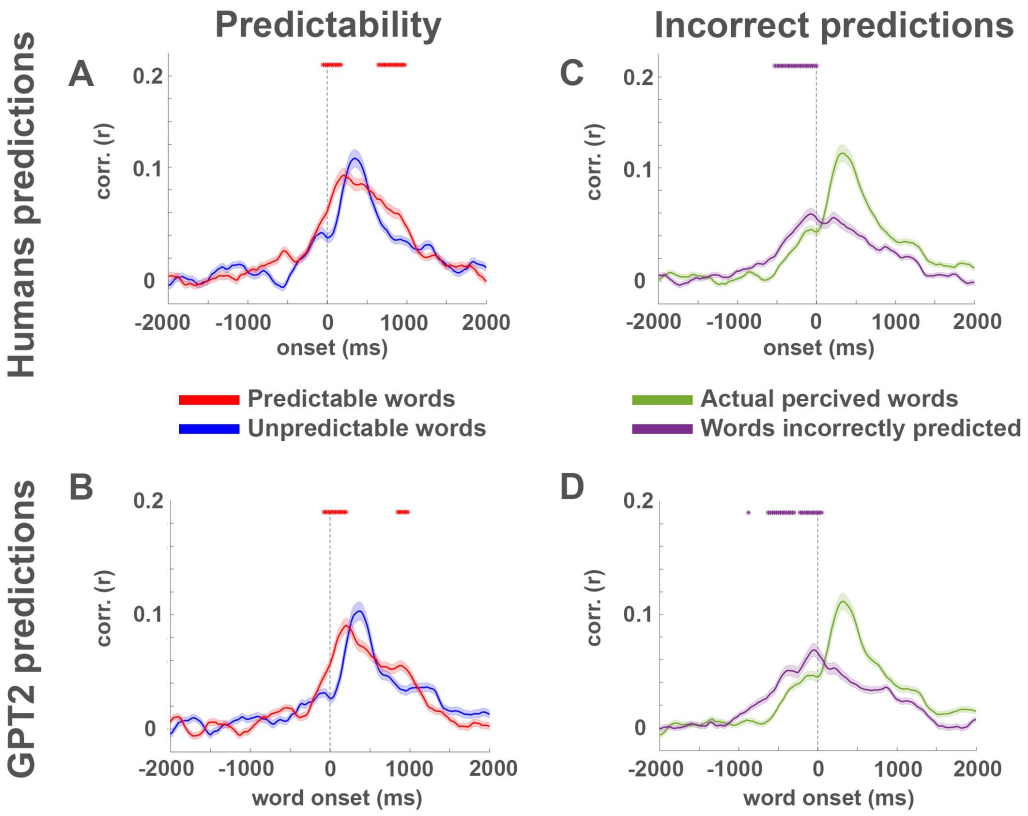

Figure 4. Modeling of neural signals before word onset for predictable, unpredictable, and incorrectly predicted words. A) Encoding model for highly predictable words (red) and unpredictable words (blue) based on human predictability scores. **B)** Same analysis as in A, using GPT2's predictability scores. **C)** Encoding model for incorrect top-1 human predictions (purple), and for the actual words listeners perceived (green).**D)** Same analysis as in A, using GPT2's top-1 incorrect predictions.

Neural signals before word onset contain information about listener expectations

Next, we examined whether the encoding model captured predictive neural responses even when the predicted word turned out to be incorrect. Focusing on the incorrect predictions allowed us to model the neural signals before and after word-onset, using the predicted word's embeddings and the actually perceived word. For that analysis, we used GloVe embeddings of 1) words humans (or GPT2) incorrectly predicted; 2) words humans (or GPT2) actually perceived.

Modeling the neural responses before word onset was significantly better while using the semantic embeddings of the words listeners predicted, even when those predictions were incorrect (Fig. 4C, purple). This demonstrates that pre-word-onset neural activity contains information about what listeners actually predicted, irrespective of what they subsequently perceived. Similar results were obtained when we used GPT2's incorrect predictions (Fig. 4D). In contrast, modeling the neural responses after word onset was significantly better while using the semantic embeddings of the actual words listeners actually perceived (Fig. 4C-D, green).

The analysis of the incorrect predictions disentangles the pre-word onset processes associated with word prediction from the post-word onset processes. It demonstrates that neural signals before word-onset contain information about listeners' internal expectations. Furthermore, it demonstrates how autoregressive DLMs' behavior can be used for modeling humans' predictions at behavioral and neural levels.

Section II: Representing contextual meaning in the brain

Using contextual embeddings to predict neural responses to natural speech
Next-word prediction's objective enables autoregressive DLMs to compress a sequence of words into a contextual embedding from which the model decodes the next word. The present results have established that the brain, similar to autoregressive DLMs, is also engaged in spontaneous next-word prediction as it listens to natural speech. Given this shared computational principle, we next investigated whether the brain, like autoregressive DLMs, compresses word sequences into contextual representation.

In natural language, each word receives its full meaning based on the identity of the preceding words [45–47]. To take an example from the stimulus, consider how the words comprising the phrase "fired off a few shots" take on a particular meaning given that the story is about photography, not firearms. Furthermore, the word "shot", like many words in natural language, can have very different meanings in different contexts, such as "drinking a shot at the bar" or "making the game-winning shot". Static word embeddings, like GloVe, assign one unique vector to the word "shot" and, as such, cannot capture the context-specific meaning of a word. In contrast, contextual embeddings assign a different embedding (vector) to every word as a function of the preceding words. Here we tested whether autoregressive DLMs that compress context into contextual embeddings provide a better cognitive model for neural activity during linguistic processing than static embeddings. To test this, we extracted the contextual embeddings from an autoregressive DLM (GPT2) for each of the words in the story. To extract the contextual embedding of a word, we provided the model with the preceding sequence of all prior words (up to 1024 words) in the podcast and extracted the activation of the top contextual embedding layer (for more details, see Materials and Methods).

Localizing neural responses to natural speech using contextual embeddings
Replacing static embeddings (GloVe) with contextual embeddings (GPT2) improved encoding model performance in predicting the neural responses to words (Fig. 5A). Encoding based on contextual embeddings resulted in statistically significant correlations in 164 electrodes in LH (and 34 in RH). 71 of these electrodes were not significantly predicted by the static embeddings (GloVe). The additional electrodes revealed by contextual embedding were mainly located in high-order language areas with long processing timescales along the inferior frontal gyrus, temporal pole, posterior superior temporal gyrus, parietal lobe, and angular gyrus [40]. In addition, there was a noticeable improvement in the contextual embeddings-based encoding model in primary and supplementary motor cortices. The improvement is seen both at the peak of the encoding and in the model's ability to predict neural responses to words up to four seconds before word-onset in the most selective electrodes (Fig. 5B). The improvement in the ability to predict neural signals to each word while relying on autoregressive DLM's contextual

embeddings was robust and apparent at the single electrode level (see Fig. S2 for a selection of electrodes). These results agree with recent studies demonstrating that contextual embeddings better model neural responses to words than static semantic embeddings [18,19,48]. Next, we asked which aspects of the contextual embedding drive the improvement in modeling the neural activity.

Modeling the context versus predicting the upcoming word

The improved ability to predict neural responses before word onset using contextual embedding can be attributed to two related factors that are absent in the static (GloVe-based) word embeddings: 1) the brain, like GPT2, may aggregate information about the preceding words in the story into contextual embeddings; and 2) GPT2 embeddings contain additional predictive information, not encoded in static embeddings, about the identity of the upcoming word in the sequence. By carefully manipulating the contextual embeddings and developing an embedding-based decoder, we show how both context and next-word prediction contribute to contextual embeddings' improved ability to model the neural responses.

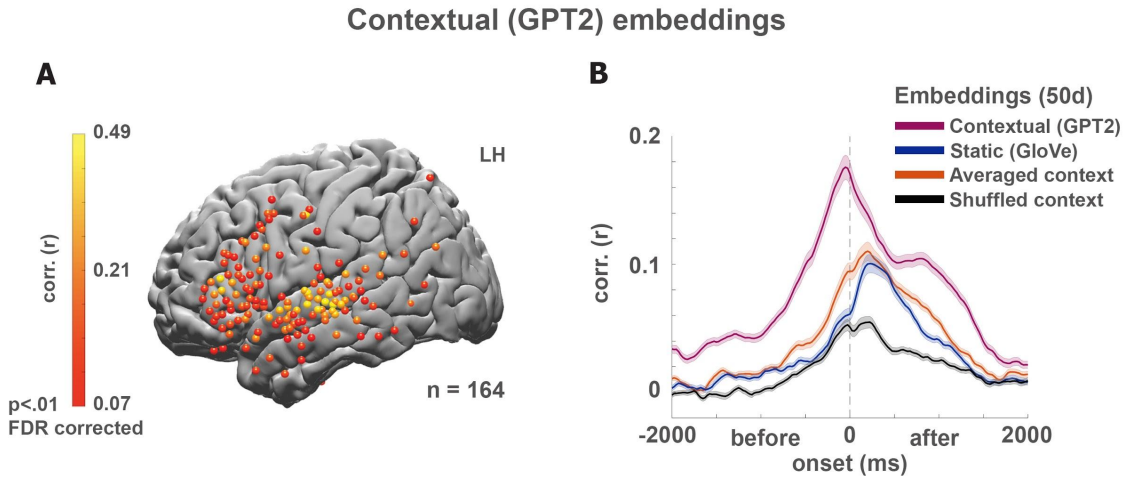

Figure 5. Contextual (GPT2) embeddings further improve the modeling of neural responses before word onset. A) Peak correlation between predicted and actual word responses for the contextual (GPT2) embeddings. Using contextual embeddings significantly improved the encoding model's ability to predict the neural signals for unseen words across many electrodes. **B)** Encoding model performance for contextual embeddings (GPT2) aggregated across all electrodes with significant encoding for GloVe (Fig. 3B): contextual embeddings (purple), static embeddings (GloVe, blue), contextual embeddings averaged across all occurrences of a given word (orange), contextual embeddings shuffled across context-specific occurrence of a given word (black).

Representing word meaning in unique contexts

GPT2's capacity for representing context captures additional information in neural responses above and beyond the information encoded in GloVe. A simple way to represent the context of prior words is to combine (e.g., concatenating) the static embeddings of the preceding sequence of words. To test this simpler representation of context, we concatenated GloVe embeddings for the six preceding words in the text into a longer "context" vector and compared

the encoding model performance to GPT2's contextual embeddings (after reducing both vectors to 50 dimensions using PCA). While the concatenated static embeddings were better in predicting the prior neural responses than the original GloVe vectors (which only capture the current word), they still underperformed GPT2's encoding before word articulation (Fig. S6). This result suggests that GPT2's contextual embeddings are better suited to compress the contextual information embedded in the neural responses than static embeddings.

A complementary way to demonstrate that contextual embeddings uncover aspects of the neural activity that static embeddings cannot capture is to remove the unique contextual information from GPT2 embeddings. We removed contextual information from GPT2's contextual embeddings by averaging all tokens' embeddings for each unique word (e.g., all occurrences of the word "monkey") into a single vector. Thus, we collapsed the contextual embedding into a static embedding in which each unique word in the story is represented by one unique vector. The resulting embeddings are still specific to the overall topic of this particular podcast (unlike GloVe). Still, they do not contain the local context for each occurrence of a given word (e.g., the context in which "*monkey*" was used in sentence 5 versus the context in which it was used in sentence 50 of the podcast). Indeed, removing context from the contextual embedding by averaging the vector for each unique word effectively reduced the encoding model's performance to that of the static GloVe embeddings (Fig. 5B, orange).

Finally, we examined how the specificity of the contextual information in the contextual embeddings improved the ability to model the neural responses to each word. To that end, we scrambled the embeddings across different occurrences of the same word in the story (e.g., switched the embedding of the word "*monkey*" in sentence 5 with the embedding for the word "*monkey*" in sentence 50). This manipulation tests whether contextual embeddings are necessary for modeling neural activity for a specific sequence of words. Scrambling the same word occurrences across contexts substantially reduced the encoding model performance (Fig. 5B, black), pointing to the contextual dependency represented in the neural signals. Taken together, these results suggest that contextual embeddings provide us with a new way to model the context-dependent neural representations of words in natural contexts.

Using contextual embeddings for predicting the next word from neural responses

The encoding results we report here provide evidence that: 1) Like autoregressive DLMs, the brain is spontaneously engaged in next-word prediction; 2) contextual embeddings derived from autoregressive DLMs provide a better model for predicting the neural activity before word onset. Finally, we apply a **decoding** analysis to demonstrate that in addition to better modeling the neural responses to context, contextual embeddings also improve our ability to read information from the neural responses as to the identity of upcoming words.

The **encoding** model finds a mapping from the embedding space to the neural responses which is used during the test phase for predicting neural responses to novel words, not seen in training. The **decoding** analysis inverts this procedure to find a mapping from neural responses, across multiple electrodes and time points, to the embedding space which is used during the test phase for predicting words' identity in new contexts [49]. This decoding analysis provides complementary insights to the encoding analysis by aggregating across electrodes to quantify

how much predictive information about each word's identity is embedded in the spatiotemporal patterns neural activity before and after word-onset.

The decoding analysis was performed in two steps. First, we trained a deep convolutional neural network to aggregate neural responses (Fig. 6A) and map this neural signal to the arbitrary, static (GloVe based) and to the contextual (GPT2 based) embedding spaces (Fig. 6B). Second, the predicted word embeddings were used for word classification based on their cosine-distance from all embeddings in the dataset (Fig. 6C). Although we evaluated the decoding model using classification, the classifier predictions were constrained to rely only on the embedding space's information. This is a more conservative approach than an end-to-end word classification, which may capitalize on acoustic information in the neural signals that are not encoded in the language models. As we focus on contextual embeddings, we used the electrodes that were significant for contextual encoding (Fig. 5A). Similar results are obtained for the choice of GloVe significant encoding (Fig. S7, for additional details see section Decoding model over time in supplementary Materials and Methods)

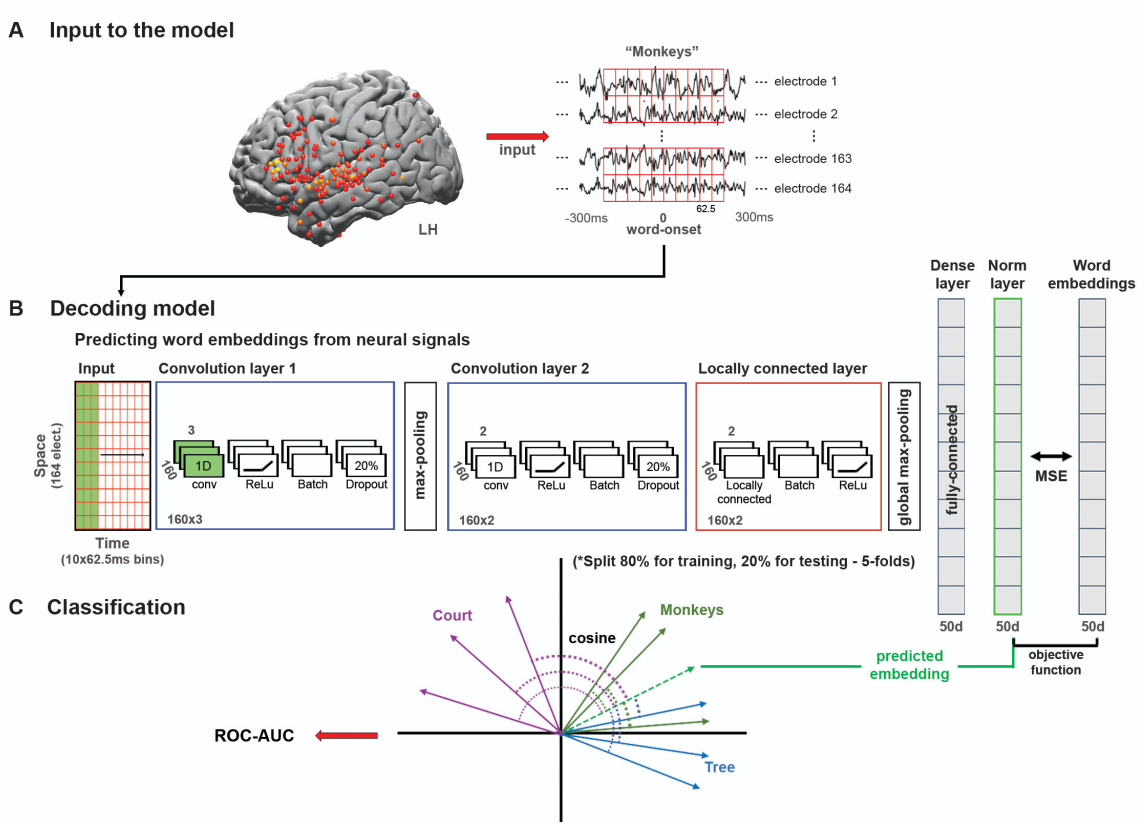

Figure 6. Deep nonlinear decoding model used to predict words from neural responses before and after word-onset. A) Neural data from left hemisphere electrodes with significant encoding model performance using GPT2 embeddings (from Fig. 5A) were used as input to the decoding model. The stimulus is segmented into individual words and aligned to the brain signal at each lag. **B)** Schematic of the feedforward deep neural network model that learns to project the neural signals for the words into the contextual embedding (GPT2) space or into the static word embedding (GloVe) space (for full description, see Appendix II). The model was trained to minimize the mean squared error (MSE) when mapping the neural signals into the embedding space. **C)** The decoding model was evaluated using a word

classification task. The quality of word classification is based on the embedding space used to construct ROC-AUC scores. This enables to assess how much information about specific words is extractible from the neural activity via the linguistic embedding space.

Using a contextual decoder greatly improved our ability to classify words' identity over decoders relying on static and arbitrary embeddings (Fig. 7). We evaluated classification performance using the area under the receiver operating characteristic curve (ROC-AUC). A model that only learns to use word frequency statistics (e.g., only guesses the most frequent word) will result in a ROC-AUC curve that falls on the diagonal line (AUC = 0.5) suggests the classifier does not discriminate between the words[50]. Classification using GPT2 (average AUC of 0.73) outperformed GloVe and arbitrary embeddings (average AUC of 0.67 and 0.63, respectively) before and after word-onset.

A closer inspection of the GPT2-based decoder indicates that the classifier managed to detect reliable information about the identity of words several hundred milliseconds before word onset (Fig. 7). In particular, starting at about -1500 ms before word onset, when the neural signals were integrated across a window of 625 ms, the classifier detected predictive information about the next word's identity. The information about the next word's identity gradually increased and peaked at an average AUC of 0.73 at a lag of 250 ms after word onset, when the signal was integrated across a window from -62.5 ms to 562.5 ms. GloVe embeddings show a similar trend with a marked reduction in classifier performance (Fig. 7, blue). The decoding model's capacity to classify words before word onset demonstrates that the neural signal contains a considerable amount of predictive information about the meaning of the next word, up to a second before it is perceived. At long time scales (more than 2 seconds), all decoders' performance dropped to chance.

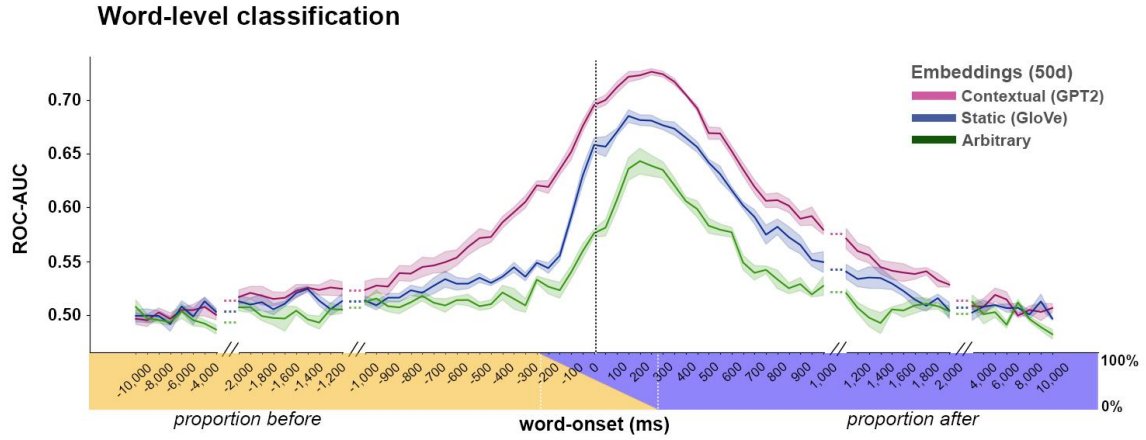

Figure 7. Using a decoding model for classification of words before and after word-onset. *Word-level classification.* Classification performance for contextual embeddings (GPT2; purple), static embeddings (GloVe; blue), and arbitrary embeddings (green). The *x*-axis labels indicate the center of each 625 ms window used for decoding at each lag (between -10 to 10 sec). The colored stripe indicates the proportion of pre- (yellow) and post- (blue) word onset time points in each lag. Error bands denote SE across five test folds. Note that contextual embeddings improve classification performance over GloVe both before and after word-onset.

Discussion

Deep language models (DLMs) provide a new modeling framework that drastically departs from classical psycholinguistic models. They are not designed to learn a concise set of interpretable syntactic rules to be implemented in novel situations, nor do they rely on part of speech concepts or other linguistic terms. Rather, they learn from surface-level linguistic behavior to predict and generate the contextually appropriate linguistic outputs. The current paper provides compelling behavioral and neural evidence for deep connections between autoregressive DLMs and the human brain.

Spontaneous prediction as a keystone of language processing

Autoregressive DLMs learn according to the simple self-supervised objective of context-based next-word prediction. The extent to which proficient English speakers are spontaneously engaged in next-word predictions as they listen to continuous, minutes-long, natural speech has been underspecified. Our behavioral results revealed a robust capacity for next-word prediction in real-life stimuli, which matches a modern autoregressive DLM (Fig. 1). Our findings demonstrate that the brain constantly and actively predicts forthcoming words during passive listening to natural speech. The predictive neural signals are robust, and can be detected hundreds of milliseconds before word-onset. Notably, the next-word prediction processes are associated with listeners' expectations and can be dissociated from the processing of the actually perceived words after word-onset (Fig. 4). The active prediction reported in this study is distinct from the neural processes observed 200–600 ms after word onset associated with prediction error and surprise (e.g., N400 and P600 event-related potentials [32–36]). Together, these findings reveal a shared computational principle between how the brain and deep autoregressive language models process incoming linguistic information in natural contexts.

Context-specific meaning as a keystone of language processing

Language is fundamentally contextual, as each word attains its full meaning in the context of preceding words over multiple timescales [40,53]. Even a single change to one word or one sentence at the beginning of a story can alter the neural responses to all subsequent sentences [46,54]. The contextual word embeddings learned by DLMs provide a new way to compress linguistic context into a numeric vector space, which outperforms the use of static semantic embeddings. While static embeddings and contextual embeddings are fundamentally different, our neural results also hint at how they relate to each other. Our results indicate that both static and contextual embeddings can predict neural responses to single words in many language areas[19] along the superior temporal cortex, parietal lobe, and inferior frontal gyrus. Switching from static to contextual embeddings boosted our ability to model neural responses during natural speech processing across many of these brain areas. Finally, averaging contextual embeddings associated with a given word, resulted in a static embedding for each word comparable to GloVe embeddings in predicting the neural responses to single words in the story. Taken together, these results suggest that the brain is coding for the semantic relationship among words contained in static embeddings while also being tuned to the unique contextual relationship between the specific word and the preceding words in the sequence[55].

Using autoregressive language model as a cognitive model

Can DLMs, such as GPT2, provide insights into the cognitive mechanisms underpinning the human language faculty. We hypothesize that the family of DLMs is sharing certain critical computational principles with biological language. This does not imply that they are identical, nor that they share the same circuit architecture[56]. Human brains and DLMs share computational principles [57,58], but they are likely to implement them using radically different neural architectures [56,59]. Many state-of-the-art DLMs rely on transformers, a type of neural network architecture developed to solve sequence transduction. While current DLMs are an impressive engineering achievement, they are not biologically feasible. They are designed to parallelize a task that is largely computed serially, word by word, in the human brain. There are many ways to transduce a sequence into a contextual embedding vector. To the extent that the brain relies on a next-word prediction objective to learn how to use language in context, it likely uses a different implementation[57].

Such a learning procedure relies on gradually exposing the model to millions of real-life examples. Our finding of spontaneous predictive neural signals as participants listen to natural speech suggests that active prediction may underlie humans' lifelong language learning. Indeed, observational work in developmental psychology suggests that children are exposed to tens of thousands of words in contextualized speech each day, creating a large data volume available for learning [60–62]. Future studies, however, will have to assess whether these cognitively plausible, prediction-based feedback signals are at the basis of human language learning and whether the brain is using such predictive signals to guide language acquisition. Finally, while predicting words across multiple timescales may be an effective learning objective for language acquisition, it is by no means the only feasible objective. Thus, it is likely that the brain relies on additional simple objectives at different timescales to facilitate learning [59,63].

Psycholinguistic versus deep language models

DLMs try to solve a fundamentally different problem than psycholinguistic language models. Psycholinguistic language models aim to uncover a set of generative (learned or innate) rules to be used in infinite, novel situations with extrapolation as the key principle needed for generating new sentences[64]. Finding a set of linguistic rules, however, was deemed challenging. There are numerous exceptions for every rule, conditioned by discourse context, meaning, dialect, genre, and many other factors. In contrast, DLMs aim to provide the appropriate linguistic output given the prior statistics of language use in similar contexts [59,65]. In other words, psycholinguistic theories aim to describe observed language in terms of a succinct set of explanatory constructs while DLMs deemphasize interpretability. Such models are performance-oriented and are focused on learning how to generate formed linguistic outputs as a function of context [66]. The reliance on performance creates an interesting connection between DLMs and usage- (context-) based constructionist approaches to language [67–69]. Furthermore, DLMs avoid the circularity built into many psycholinguistic language models that rely on linguistic terms to explain how language is encoded in neural substrates [20,70]. Finally, the internal contextual embedding space in DLMs can capture many aspects of the latent structure of human language, including syntactic trees, voice, co-references, morphology, and long-range semantic and pragmatic dependencies [1,71–73]. Such findings demonstrate the power of brute-force memorization and

interpolation in learning how to generate the appropriate linguistic outputs in light of prior contexts[59].

Conclusion

Linguistics aims to expose the hidden underlying structure of language. This paper provides evidence for shared core computational principles of next-word prediction in context between deep language models and the human brain. While DLMs may provide a building block for our high-level cognitive faculties, they undeniably lack certain central hallmarks of human cognition. Linguists were primarily interested in how we construct well-formed sentences, exemplified by the famous grammatically correct but meaningless sentence composed by Noam Chomsky "colorless green ideas sleep furiously"[2]. Similarly, DLMs are generative in the narrow linguistic sense of being able to generate new sentences that are grammatically, semantically, and even pragmatically well-formed at a superficial level. However, although language may play a central organizing role in our cognition, linguistic competence is insufficient to capture thinking. Unlike humans, DLMs cannot think, understand, or generate new meaningful ideas by integrating prior knowledge. They simply echo the statistics of their input[74]. A core question for future studies in cognitive neuroscience and machine learning is how the brain can leverage predictive, contextualized linguistic representations, like those learned by DLMs, as a substrate for generating and articulating new thoughts.

Acknowledgements

We thank Adele Goldberg, Rita Goldstein, Sebastian Michelmann, Meir Meshulam, Manoj Kumar, Malcolm Slaney for technical and conceptual assistance that motivated and informed this manuscript's writing. This work was supported by the National Institutes of Health under award numbers DP1HD091948 (A.G, Z.Z,, A.P, B.A, G.C, A.R, C.K, F.L, A.F and U.H.), R01MH112566 (S.A.N.), NIH R01NS109367-01 to A.F, Finding A Cure for Epilepsy and Seizures (FACES), and DataX Fund, Schmidt Futures Foundation.

Materials and Methods

Transcription and alignment

Stimuli for the behavioral test and ECoG experiment were extracted from a 30-minute story "So a Monkey and a Horse Walk Into a Bar: Act One, Monkey in the Middle" taken from the This American Life podcast. The story was manually transcribed and aligned to the audio by marking the onset and offset of each word. Sounds such as laughter, breathing, lip-smacking, applause, and silent periods were also marked to improve the alignment's accuracy. The audio was downsampled to 11 kHz and the Penn Phonetics Lab Forced Aligner was used to automatically align the audio to the transcript[75]. The forced aligner uses a phonetic Hidden Markov model to find the temporal onset and offset of each word and phoneme in the story. After automatic alignment was complete, the alignment was manually evaluated by an independent listener..

Behavioral word-prediction experiment

To obtain a continuous measure of prediction, we developed a novel sliding-window behavioral paradigm where healthy adult participants made predictions for each upcoming word in the story. 300 participants completed a behavioral experiment on Mechanical Turk. Since predicting each word in a 30-minute (5113 words) story is taxing, we divided the story into six segments and recruited six non-overlapping groups of 50 participants to predict every upcoming word within each segment of the story (about 830 words per group of participants). The first group was exposed to the first two words in the story and then asked to predict the upcoming (i.e., third) word. After entering their prediction, the actual next word was revealed, and participants were asked again to predict the next upcoming (i.e., fourth) word in the story. Once 10 words were displayed on the screen, the left-most word was removed and the next word was presented (Fig. 1B). The procedure was repeated, using a sliding window until the first group provided predictions for each word in the story's first segment. Each of the other five groups listened uninterruptedly to the prior segments of the narrative and started to predict the next word at the beginning of their assigned segments.

Next, we calculated a mean prediction performance (proportion of participants predicting the correct word) across all 50 listeners for each word in the narrative, which we refer to as the "predictability score" (Fig. 1C). A predictability score of 1 indicates that all subjects correctly guessed the next word, and a predictability score of 0 indicates that no participant predicted the upcoming word. Due to a technical error, data for 33 words were omitted, and thus the final data contained 5078 words.

ECoG experiment

Nine patients (5 female; 20–48 years old) listened to the same story stimulus from beginning to end. Participants were not explicitly made aware that we would be examining word prediction in our subsequent analyses. One patient was removed due to excessive epileptic activity and low SNR across all experimental data collected during the day. All patients experienced pharmacologically refractory complex partial seizures and volunteered for this study via the New York University School of Medicine Comprehensive Epilepsy Center. All participants had elected to undergo intracranial monitoring for clinical purposes and provided oral and written informed consent before study participation, according to the New York University Langone Medical Center Institutional Review Board. Patients were informed that participation in the study was

unrelated to their clinical care and that they could withdraw from the study at any point without affecting their medical treatment.

For each patient, electrode placement was determined by clinicians based on clinical criteria (Fig. 2A). One patient was consented to have an FDA-approved hybrid clinical-research grid implanted which includes standard clinical electrodes as well as additional electrodes in between clinical contacts. The hybrid grid provides a higher spatial coverage without changing clinical acquisition or grid placement. 917 electrodes were placed on the left hemisphere and 233 on the right hemisphere. Brain activity was recorded from a total of 1086 intracranially implanted subdural platinum-iridium electrodes embedded in silastic sheets (2.3 mm diameter contacts, Ad-Tech Medical Instrument, for the hybrid grids 64 standard contacts were 2 mm diameter and additional 64 contacts were 1 mm diameter, PMT corporation, Chanassen, MN). Decisions related to electrode placement and invasive monitoring duration were determined solely on clinical grounds without reference to this or any other research study. Electrodes were arranged as grid arrays (8 × 8 contacts, 10 or 5 mm center-to-center spacing), or linear strips. Altogether, the subdural electrodes covered extensive portions of lateral frontal, parietal, occipital, and temporal cortex of the left and/or right hemisphere (Fig. 3A for electrode coverage across all subjects).

Recordings from grid, strip and depth electrode arrays were acquired using one of two amplifier types: NicoletOne C64 clinical amplifier (Natus Neurologics, Middleton, WI), bandpass filtered from 0.16–250 Hz, and digitized at 512 Hz; Neuroworks Quantum Amplifier (Natus Biomedical, Appleton, WI) recorded at 2048 Hz, highpass filtered at 0.01 Hz and then resampled to 512 Hz. Intracranial EEG signals were referenced to a two-contact subdural strip facing towards the skull near the craniotomy site. All electrodes were visually inspected, and those with excessive noise artifacts, epileptiform activity, excessive noise, or no signal were removed from subsequent analysis (164/1065 electrodes removed).

Pre-surgical and post-surgical T1-weighted MRIs were acquired for each patient, and the location of the electrodes relative to the cortical surface was determined from co-registered MRIs or CTs following the procedure described by Yang and colleagues[76]. Co-registered, skull-stripped T1 images were nonlinearly registered to an MNI152 template and electrode locations were then extracted in Montreal Neurological Institute (MNI) space (projected to the surface) using the co-registered image. All electrode maps are displayed on a surface plot of the template, using the Electrode Localization Toolbox for MATLAB available at (https://github.com/HughWXY/ntools_elec).

Preprocessing

Data analysis was performed using the FieldTrip toolbox[77], along with custom preprocessing scripts written in MATLAB 2019a (MathWorks). The time course of signal power was estimated using Morlet wavelets. The power time course was computed in the frequency range of 70-200Hz separately for each frequency in steps of 5Hz. We excluded harmonics of line noise of 120 and 180 Hz, and used the logarithm of each power time course estimate. These estimates were z-scored, and then averaged across these frequencies to create a high gamma

band time course. This broadband power time course was then smoothed with a 50 ms Hamming window.

Large spikes exceeding 4 quartiles above and below the median were removed and replacement samples were imputed using cubic interpolation. We then re-referenced the data to account for shared signals across all channels using either the Common Average Referencing (CAR) method[77,78] or an ICA-based method[79] (based on the participant's noise profile). High-frequency broadband (HFBB) power frequency provided evidence for a high positive correlation between local neural firing rates and high gamma activity[80]. The high gamma band fluctuation exhibited good estimations in the neural spiking population near each electrode[81]. After computing the broadband power time course, the power estimates were divided by the mean value. This method improves the signal-to-noise ratio in the estimate of high-frequency power[82].

Encoding analysis

In this analysis, a linear model is implemented for each lag for each electrode relative to word-onset, and is used to predict the neural signal from word embeddings (Fig. 2B). The values calculated are the correlations between the predicted signal and held out actual signal at each lag (separately for each electrode), indicating the linear model's performance. Before fitting the linear models for each time point, we implement a running window averaging across a 200 ms window. We assess the linear models' performance (model for each lag) in predicting neural responses for held-out data using a 10-fold cross-validation procedure. The neural data were randomly split into a training set (i.e., 90% of the words) for model training and a testing set (i.e., 10% of the words) for model validation. On each fold of this cross-validation procedure, we used ordinary least-squares multiple linear regression to estimate the regression weights from 90% of the words. We then applied those weights to predict the neural responses to the other 10% of the words. The predicted responses for all ten folds were concatenated so a correlation between the predicted signal and actual signal was computed over all the words of the story. This entire procedure was repeated at 161 lags from -2000 ms to 2000 ms in 25 ms increments relative to word onset.

Significance tests

To identify significant electrodes, we used a randomization procedure. At each iteration, we randomized each electrode's signal phase, thus disconnecting the relationship between the words and the brain signal but preserving the autocorrelation in the signal. Then we performed the entire encoding procedure for each electrode. We repeated this process 5000 times. After each iteration, the encoding model's maximal value across all 161 lags was retained for each electrode. We then took the maximum value for each permutation across electrodes. This resulted in a distribution of 5000 values, which was used to determine significance for all electrodes. For each electrode a *p*-value (Fig. 3A-B, 5A) was computed as the percentile of the non-permuted encoding model's maximum value across all lags from the null distribution of 5000 maximum values. Performing a significance test using this randomization procedure evaluates the null hypothesis that there is no systematic relationship between the brain signal and the corresponding word embedding. This procedure yielded a *p*-value per electrode. Electrodes with *p*-values less than .01 were considered significant. To correct for multiple

electrodes we used false-detection-rate (FDR[83]). To statistically assess the difference between the performance of two encoding models for the same electrode at specific lags (Fig. 3C-D, 5A), we subtracted the two encoding models' values for the permuted labels. This yielded a distribution of 5000 values for each lag. Using the relevant distribution, each lag was assigned with a p-value. We adjusted the resulting p-values to control the false discovery rate[83].

To test each lag's significance for the average encoding plots (Figs. 3 and 5, S2 and S5), we used a bootstrap hypothesis test to compute a *p*-value for each lag[84]. For each bootstrap, a sample matching the subset size was drawn with replacement from the encoding performance values for the subset of electrodes. The mean of each bootstrap sample was computed. This resulted in a bootstrap distribution of 5000 mean performance values for each lag. The bootstrap distribution was then shifted by the observed value to perform a null hypothesis[84]. To account for multiple tests across lags, we adjusted the resulting *p*-values to control the false discovery rate[FDR; 83]. A threshold was chosen to control the FDR at .01. We used a permutation test to assess significant differences in the average encoding (Fig. 4C-D, S3, and S4): we randomly swapped the assignment of the encoding performance values between the two models at each lag (50% of the pairs were swapped). Then we computed the average of the pairwise differences to generate a null distribution at each lag. We then calculated a *p*-value for each lag, which was corrected for multiple comparisons using FDR.

Contextual embedding extraction

We extracted contextualized word embeddings from GPT2 for our analysis. To do so, we first converted our words to tokens which were either whole words or sub-words (there's -> there 's). We only used words that were the same before and after tokenization to keep the word alignment intact. We used a sliding window of 1024 tokens, moving one token at a time, to extract the embedding for the final word in the sequence (i.e. the word and its history). Encoding these tokens into integer labels, we then put them into the model, and in return, we received the activations at each layer in the network (also known as a hidden state). GPT-2 has 48, but we focus only on the final one, before the classification layer. Finally, the token of interest was the final word of the sequence, yet we used the second-to-last token as the hidden state for the last word because it was the same activation embedding that was used to predict that word.

Decoding analysis

The goal of this analysis was to predict words from the neural signal. The input neural data were averaged in 10 62.5-ms bins spanning 625 ms for each lag. Each bin consisted of 32 data points (the neural recording sampling rate was 512Hz).

The neural network decoder (see architecture in Appendix II) was trained to predict a word's embedding from the neural signal at a specific lag. The data was split into 5 non-overlapping temporal intervals (i.e., folds) and used in a cross-validation procedure. Each fold consisted of a mean of 717.04 training words (SD = 1.32). Three folds were used for training the decoder (training set), one fold was used for early stopping (development set), and one fold was used to assess model generalization (test set). The neural net was optimized to minimize the MSE when predicting the embedding. The decoding performance was evaluated using a classification task

assessing how well the decoder can predict the word label from the neural signal. We used the receiver operating characteristic curve (ROC-AUC) measure.

To calculate the ROC-AUC, we computed the cosine distance between each of the predicted embeddings and the embeddings of all instances of each unique word label. The distances were averaged across unique word labels, yielding one score for each word label (i.e., logit). We used a softmax transformation on these scores (logits). For each label (classifier), we used the logits and the information of whether the instance matched the label to compute a ROC-AUC for the label. We plotted the weighted ROC-AUC according to the word's frequency in the test set (which was equal to the frequency in the training set due to the stratified split). We chose words with at least 5 repetitions in the training set (69% of the overall words in the narrative; see Appendix I for word list).

To improve the performance of the decoder, we implemented an ensemble of models. We independently trained 10 decoders with randomized weight initializations and randomized the batch order fed into the neural net for each lag. This procedure generated 10 predicted embeddings. Thus, for each predicted embedding, we repeated the distance calculation from each word label 10 times. These 10 values were averaged and later used for ROC-AUC.

Supplementary Information

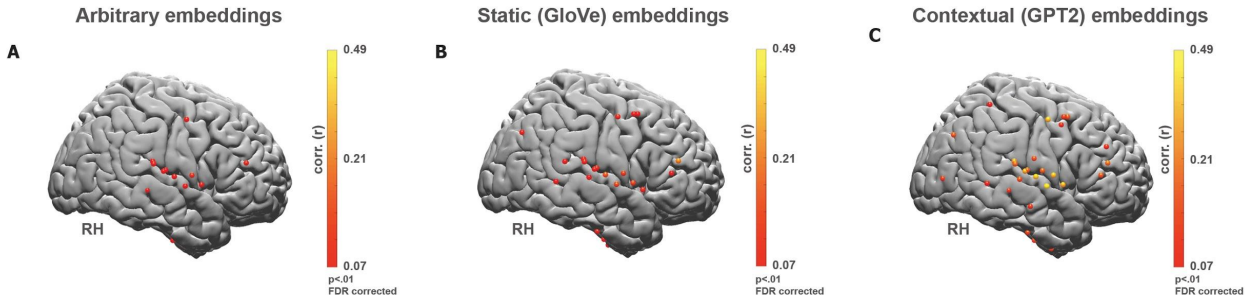

Figure S1. Right hemisphere encoding results show an advantage for contextual (GPT2) embeddings over static (GloVe) and arbitrary embeddings. Right Hemisphere maps for correlation between. **A)** Predicted and actual word responses for the arbitrary embeddings (nonparametric permutation test; $p < .01$, FDR corrected). **B)** Correlation between predicted and actual word responses for the static (GloVe) embeddings. **C)** Correlation between predicted and actual word responses for the contextual (GPT2) embeddings. Using contextual embeddings significantly improved the encoding model's ability to predict the neural signals for unseen words across many electrodes. Given that we had fewer electrodes in the right hemisphere relative to the left hemisphere, this study is not set up to test differences in language lateralization across hemispheres.

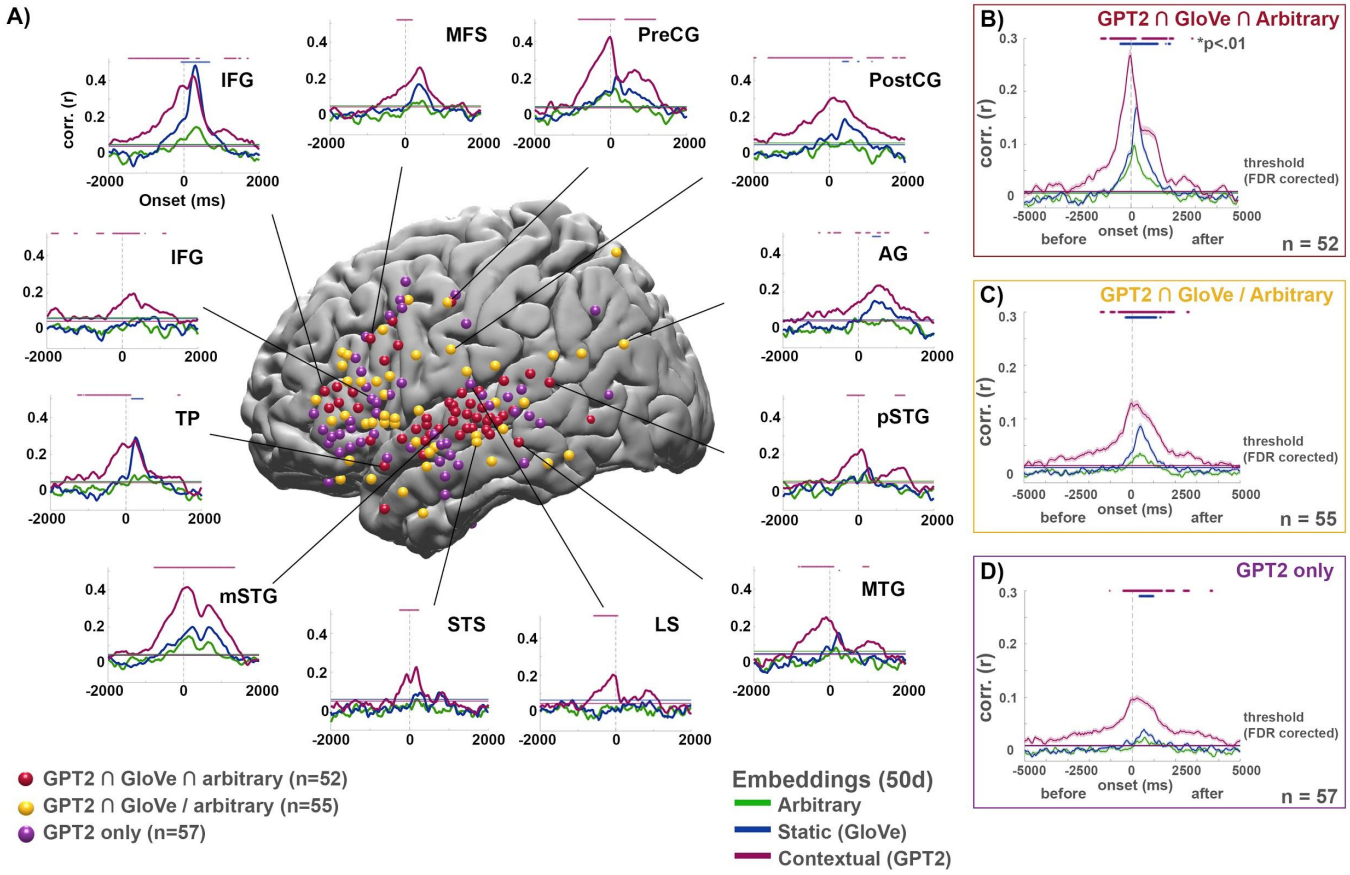

Figure S2. Contextual embedding significantly improves the modeling of neural signals. A) Map of the electrodes in the left hemisphere with significant encoding for 1) all three types of embeddings (GPT2 ∩ GloVe ∩ arbitrary, red); 2) for static and contextual embeddings (GPT2 ∩ GloVe, but not arbitrary, yellow); 3) and contextual only (GPT2, purple) embeddings. Note the three groups do not overlap. A sampling of encoding performance for selected individual electrodes across different brain areas: inferior frontal gyrus (IFG), temporal pole (TP), medial superior central gyrus (mSTG), superior temporal sulcus (STS), lateral sulcus (LS), middle temporal gyrus (MTG), posterior superior temporal gyrus (pSTG), angular gyrus (AG), post central gyrus (postCG), precentral gyrus (PreCG), and middle frontal sulcus (MFS). (Green - encoding for the arbitrary embeddings, blue - encoding for static (GloVe) embeddings; purple - encoding for contextual (GPT2) embeddings. **B)** Average encoding model performance across lags for all electrodes with significant encoding for the three types of encoding (52 electrodes marked in red). **C)** Average encoding model performance across lags for all electrodes with significant encoding only for the GloVe and GPT2, but not arbitrary (55 electrodes marked in yellow). **D)** Average encoding model performance across lags for all electrodes with significant encoding for GPT2-only (57 electrodes marked in purple). The lines indicate average performance at each lag relative to word onset, the standard error bands indicate standard error of the encoding model across electrodes. The horizontal lines specify the statistical threshold after correcting for multiple comparisons ($p < .01$, FDR). Blue asterisks indicate lags for which GloVe embeddings significantly outperform arbitrary embeddings ($p < .01$), and purple asterisks indicate lags for which GPT2 embeddings significantly outperform GloVe embeddings ($p < .01$, nonparametric permutation test, FDR corrected).

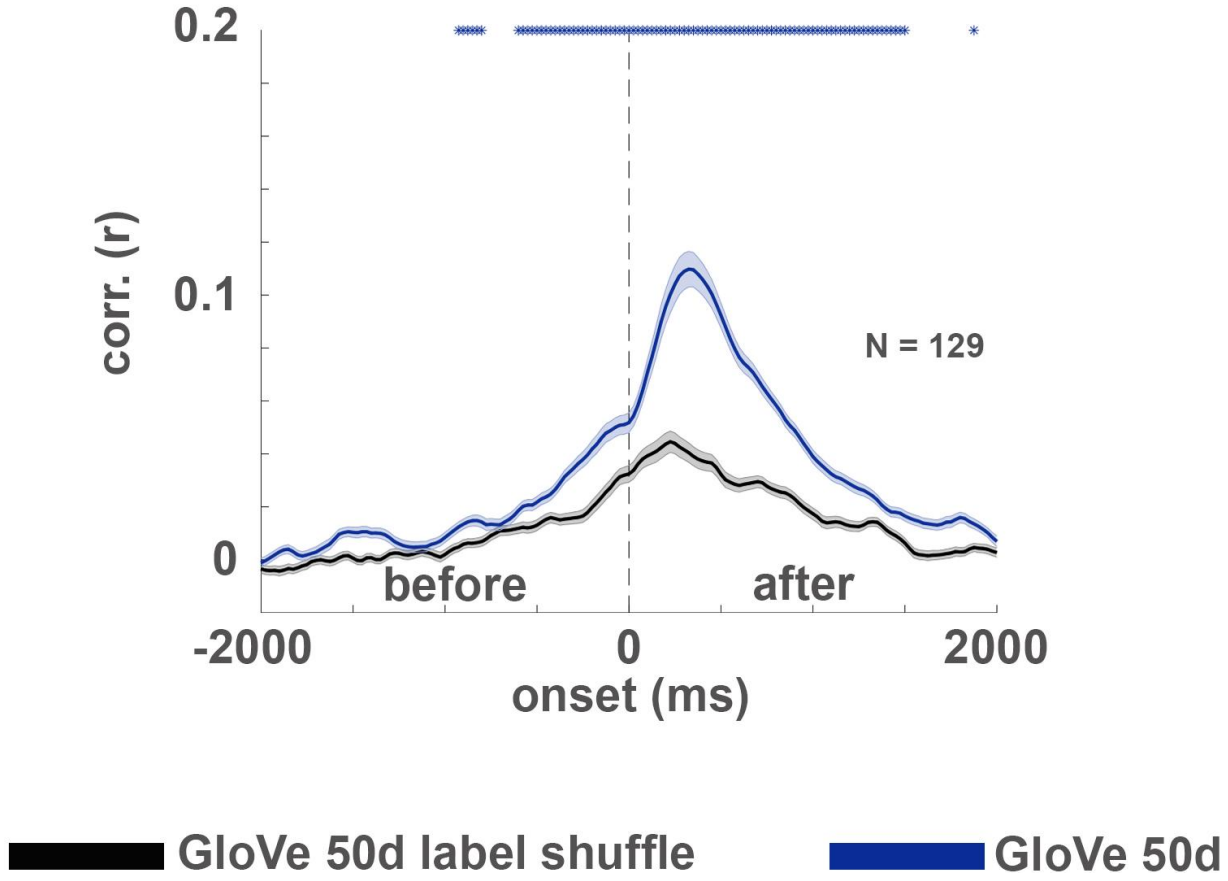

Figure S3. GloVe's space embedding attributes. It can be argued that GloVe based encoding outperforms arbitrary-based encoding due to a general property of the space that GloVe embeddings induce (for example, they are closer / further away from each other). To control for this possible confound, we consistently mismatched the labels of the embeddings of GloVe and used the mismatched version for encoding. This means that each unique word was consistently matched with a specific vector that is actually an embedding of a different label (for example, matching each instance of the word 'David' with the embedding of the word 'court'). This manipulation uses the same embedding space that GloVe uses and also induces a consistent mapping of words to embeddings (as in the arbitrary-based encoding). The matched GloVe (blue) outperformed the mismatched GloVe (black), supporting the claim that GloVe embedding carries information about words statistics that is useful for predicting the brain signal.

-

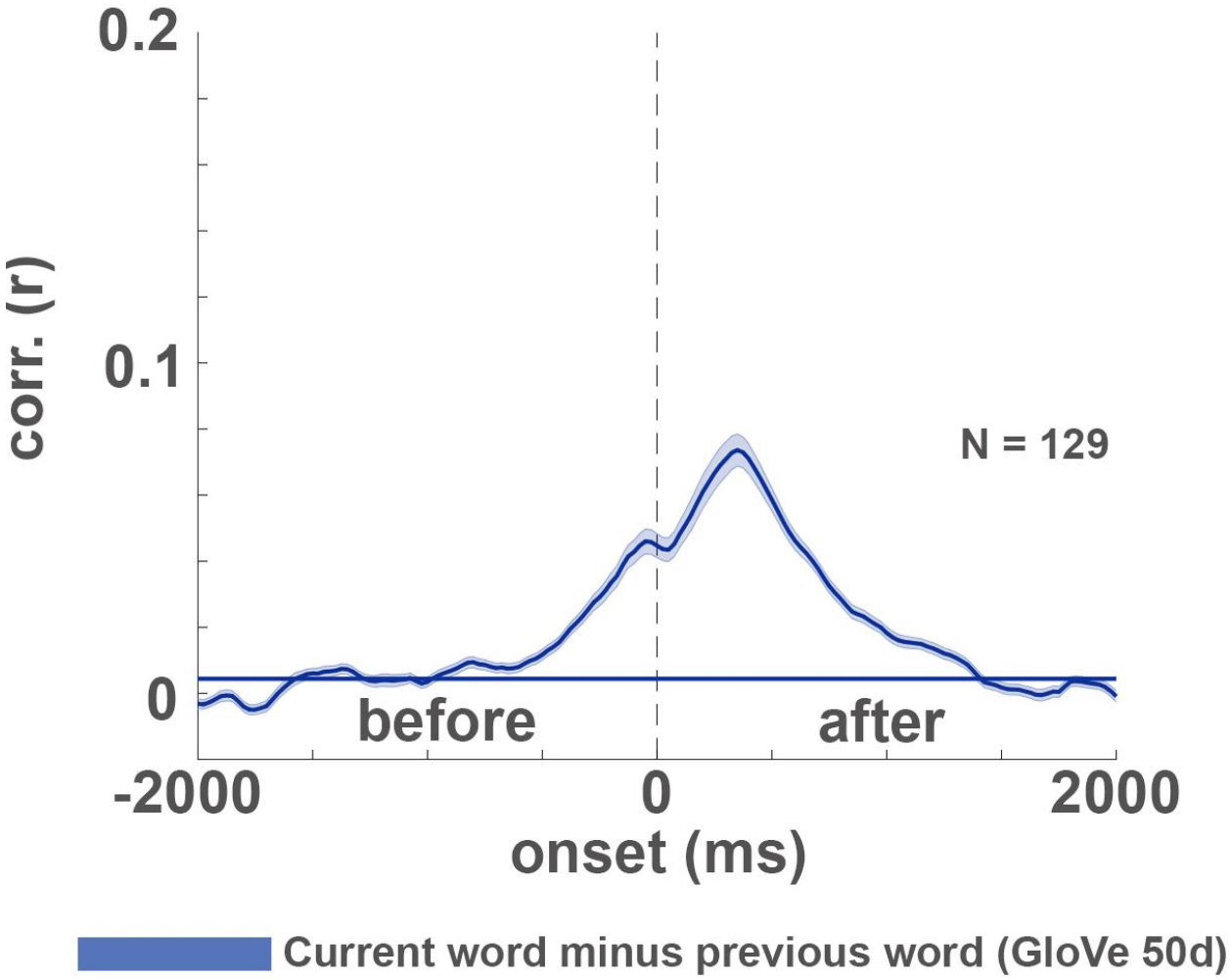

Figure S4. Controlling co-occurrence-induced correlations in GloVe. When focusing on the predictive signal (i.e., correlations before the onset), one may suggest that the predictive encoding is stemming from co-occurrences of words (bigrams) if two words occur (or if their embedding correlates with each other), the correlations before onset may reflect the relation between the labels or their embeddings but not a correlation between the current embedding and the neural signal that preceded it. To ensure that the signal predicted before onset is not a result of an indirect correlation, we regressed out the embedding of the previous word from each word and re-ran the encoding analysis. This also yielded a significant encoding model before and after word-onset. This indicates that the encoding before onset is not the result of a correlation between adjacent embeddings or the words' co-occurrence.

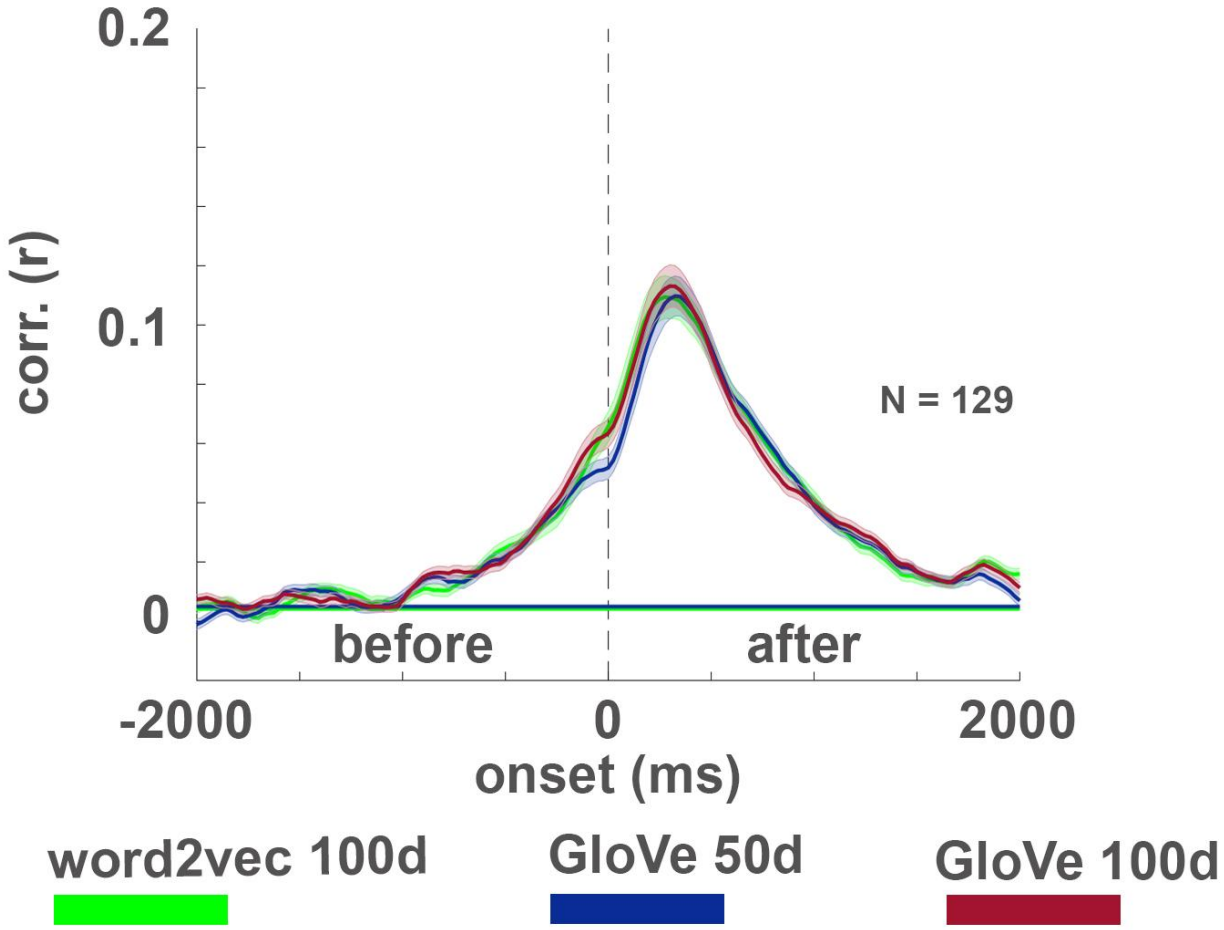

Figure S5. Comparison of GloVe- and word2vec-based static embeddings. The encoding procedure was repeated for two additional static embeddings using the electrodes that were found significant for GloVe-50 encoding on the left hemisphere (Fig. 3B). Each line indicates the encoding model performance averaged across electrodes for a given type of static embedding at lags from -2000 to 2000 ms relative to word onset. The error bands indicate the standard error of the mean across the electrodes at each lag. 100-dimensional word2vec and GloVe embeddings resulted in similar encoding results to the initial 50-dimensional GloVe embeddings. This suggests that results obtained with static embeddings are robust to the specific type of static embeddings used.

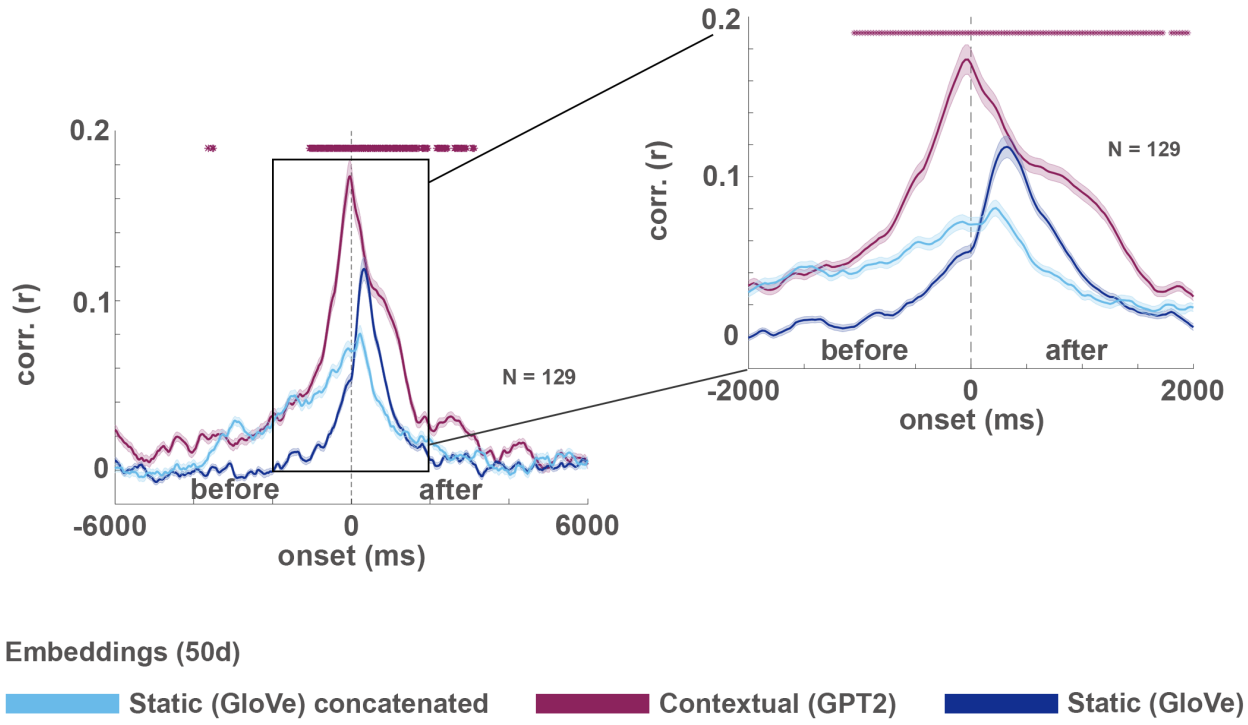

Figure S6. Comparison of GPT2 and concatenation of static embeddings. The increased performance of GPT2 based contextual embeddings encoding may be attributed to the fact that it consists of information about the previous words' identity. To examine this possibility, we concatenated 5 GloVe words and reduced their dimensionality to 50 features. GPT2 based encoding outperformed the mere concatenation before word onset, suggesting that GPT2's ability to compress the contextual information improves the ability to model the neural signals before word onset.

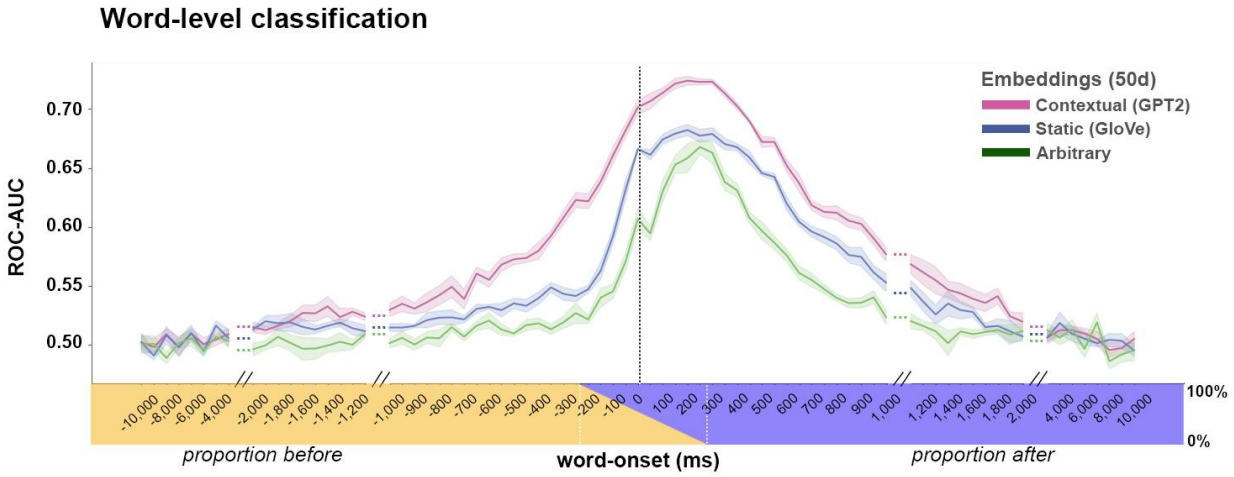

Figure S7. Decoding model using GloVe significant electrodes as input. *Word-level classification* for contextual embeddings (GPT2; purple), static embeddings (GloVe; blue), and arbitrary embeddings (green). The *x*-axis labels indicate the center of each 625 ms window used for decoding at each lag (between -10 to 10 sec). The colored strip indicates the proportion of pre- (yellow) and post- (blue) word onset time points in each lag. Error bands denote SE across five test folds. Note that contextual embeddings improve classification performance over GloVe both before and after word-onset.

Appendix I - Word List

a	called	have	make	property	they	which
about	camera	he	me	public	think	who
after	case	him	monkey	really	this	wikipedia
all	copyright	his	my	right	thought	with
an	could	how	next	said	to	would
and	court	human	no	saw	twenty	yeah
andrew	david	i	not	say	two	year
animal	day	if	now	see	uh	you
are	did	in	of	shot	um	your
argument	do	into	on	should	up	
around	domain	is	one	so	very	
at	even	it	or	sued	wa	
attorney	first	judge	other	take	wales	
be	for	just	out	that	way	
because	friend	know	over	the	we	
been	from	law	own	their	well	
before	get	lawyer	people	them	were	
being	got	legal	photo	then	what	
but	ha	like	photograph	there	when	
by	had	look	picture	these	where	

Appendix II - Model Details

Layer (type)	**Output Shape**	**Param #**
input_1 (InputLayer)	[(10, 164)]	0
conv1d (Conv1D)	(8, 160)	78720
activation (Activation)	(8, 160)	0
batch_normalization (BatchNo)	(8, 160)	640
dropout (Dropout)	(8, 160)	0
max_pooling1d (MaxPooling1D)	(4, 160)	0
conv1d_1 (Conv1D)	(3, 160)	51200
activation_1 (Activation)	(3, 160)	0
batch_normalization_1 (BatchNo)	(3, 160)	640
dropout_1 (Dropout)	(3, 160)	0
Llocally_connected1d	(2, 160)	102720
batch_normalization_2 (BatchNo)	(2, 160)	640
activation_2 (Activation)	(2, 160)	0
global_max_pooling1d	(60)	0
dense (Dense)	(50)	8050
layer_normalization (LayerNo)	(50)	100
Total parameters		**242,710**
Trainable parameters		**241,750**
Non-trainable parameters		**960**

- Learning rate: 0.00025
- Batch size: 256
- Convolutional layers L2 regularization alpha: 0.003

- Dense layer L2 regularization alpha: 0.0005
- Dropout probability is 21%
- Weights averaged over last 20 epochs before early stopping
- Trained for a maximum of 1500 epochs with patience of 150 epochs

We used hyperparameter search to choose depth, batch size, learning rate, patience, convolutional filter. [85]

References

1. Linzen, T. & Baroni, M. Syntactic Structure from Deep Learning. *Annu. Rev. Linguist.* (2021) doi:10.1146/annurev-linguistics-032020-051035.
2. Chomsky, N. Syntactic Structures. (1957) doi:10.1515/9783112316009.
3. Jacobs, R. A. & Rosenbaum, P. S. *English transformational grammar*. (1968).
4. Lewis, M. *et al.* BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv [cs.CL]* (2019).
5. Yang, Z. *et al.* XLNet: Generalized Autoregressive Pretraining for Language Understanding. in *Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) 5753–5763 (Curran Associates, Inc., 2019).
6. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. (2018).
7. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* **1**, (2019).
8. Rosset, C. Turing-nlg: A 17-billion-parameter language model by microsoft. *Microsoft Blog* (2019).
9. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv [cs.CL]* (2020).
10. Cho, W. S. *et al.* Towards Coherent and Cohesive Long-form Text Generation. *Proceedings of the First Workshop on Narrative Understanding* (2019) doi:10.18653/v1/w19-2401.
11. Liu, Q., Kusner, M. J. & Blunsom, P. A Survey on Contextual Embeddings. *arXiv [cs.CL]* (2020).
12. de Vries, W. & Nissim, M. As good as new. How to successfully recycle English GPT-2 to make models for other languages. *arXiv [cs.CL]* (2020).
13. Pereira, F. *et al.* Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).
14. Makin, J. G., Moses, D. A. & Chang, E. F. Machine translation of cortical activity to text with

an encoder–decoder framework. *Nature Neuroscience* vol. 23 575–582 (2020).

15. Schwartz, D., Toneva, M. & Wehbe, L. Inducing brain-relevant bias in natural language processing models. in *Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) 14123–14133 (Curran Associates, Inc., 2019).
16. Gauthier, J. & Levy, R. Linking artificial and human neural representations of language. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019) doi:10.18653/v1/d19-1050.
17. Donhauser, P. W. & Baillet, S. Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron* **105**, 385–393.e9 (2020).
18. Jain, S. & Huth, A. G. Incorporating Context into Language Encoding Models for fMRI. doi:10.1101/327601.
19. Schrimpf, M. *et al.* The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. doi:10.1101/2020.06.26.174482.
20. McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J. & Schütze, H. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proc. Natl. Acad. Sci. U. S. A.* (2020) doi:10.1073/pnas.1910416117.
21. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & de Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. doi:10.1101/2020.12.03.410399.
22. Huang, Y. & Rao, R. P. N. Predictive coding. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 580–593 (2011).
23. Lupyan, G. & Clark, A. Words and the World: Predictive Coding and the Language-Perception-Cognition Interface. *Curr. Dir. Psychol. Sci.* **24**, 279–284 (2015).
24. Barron, H. C., Auksztulewicz, R. & Friston, K. Prediction and memory: A predictive coding account. *Prog. Neurobiol.* **192**, 101821 (2020).

25. Goldstein, A., Rivlin, I., Goldstein, A., Pertzov, Y. & Hassin, R. R. Predictions from masked motion with and without obstacles. *PLoS One* **15**, e0239839 (2020).
26. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* vol. 36 181–204 (2013).
27. Taylor, W. L. 'Cloze Procedure': A New Tool for Measuring Readability. *Journal. Q.* **30**, 415–433 (1953).
28. Kliegl, R., Nuthmann, A. & Engbert, R. Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *J. Exp. Psychol. Gen.* **135**, 12–35 (2006).
29. Fernández, G., Shalom, D. E., Kliegl, R. & Sigman, M. Eye movements during reading proverbs and regular sentences: the incoming word predictability effect. *Language, Cognition and Neuroscience* vol. 29 260–273 (2014).
30. Staub, A., Grant, M., Astheimer, L. & Cohen, A. The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language* vol. 82 1–17 (2015).
31. Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K. & Kliegl, R. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behav. Res. Methods* **51**, 1161–1178 (2019).
32. Hagoort, P., Brown, C. & Groothusen, J. The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and Cognitive Processes* vol. 8 439–483 (1993).
33. Beim Graben, P., Gerth, S. & Vasishth, S. Towards dynamical system models of language-related brain potentials. *Cogn. Neurodyn.* **2**, 229–255 (2008).
34. Kutas, M. & Federmeier, K. D. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology* vol. 62 621–647 (2011).

35. Kutas, M. & Hillyard, S. A. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* **207**, 203–205 (1980).

36. Kutas, M. & Hillyard, S. A. Brain potentials during reading reflect word expectancy and semantic association. *Nature* **307**, 161–163 (1984).

37. DeLong, K. A. & Kutas, M. Comprehending surprising sentences: sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience* **35**, 1044–1063 (2020).

38. Musiolek, L., Blankenburg, F., Ostwald, D. & Rabovsky, M. Modeling the N400 brain potential as Semantic Bayesian Surprise. in *2019 Conference on Cognitive Computational Neuroscience* (2019).

39. Chivvis & Dana. "So a Monkey and a Horse Walk Into a Bar". (2017).

40. Hasson, U., Chen, J. & Honey, C. J. Hierarchical process memory: memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–313 (2015).

41. Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* 1532–1543 (aclweb.org, 2014).

42. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. in *Advances in Neural Information Processing Systems 26* (eds. Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q.) 3111–3119 (Curran Associates, Inc., 2013).

43. Oota, S. R., Manwani, N. & Bapi, R. S. fMRI Semantic Category Decoding Using Linguistic Encoding of Word Embeddings. *Neural Information Processing* 3–15 (2018) doi:10.1007/978-3-030-04182-3_1.

44. Abnar, S., Ahmed, R., Mijnheer, M. & Zuidema, W. Experiential, Distributional and Dependency-based Word Embeddings have Complementary Roles in Decoding Brain Activity. *arXiv [cs.CL]* (2017).

45. Chen, J., Hasson, U. & Honey, C. J. Processing Timescales as an Organizing Principle for Primate Cortex. *Neuron* vol. 88 244–246 (2015).

46. Yeshurun, Y., Nguyen, M. & Hasson, U. Amplification of local changes along the timescale processing hierarchy. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9475–9480 (2017).

47. Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550 (2008).

48. Wehbe, L., Vaswani, A., Knight, K. & Mitchell, T. 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)* 233–243 (Association for Computational Linguistics, 2014).

49. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *Neuroimage* **56**, 400–410 (2011).

50. Mandrekar, J. N. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology* vol. 5 1315–1316 (2010).

51. Caucheteux, C. & King, J. R. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv* (2020).

52. Shain, C., Blank, I. A., van Schijndel, M., Schuler, W. & Fedorenko, E. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia* **138**, 107307 (2020).

53. Goldberg, A. E. *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. (Princeton University Press, 2019).

54. Yeshurun, Y. *et al.* Same Story, Different Story: The Neural Representation of Interpretive Frameworks. *Psychol. Sci.* **28**, 307–319 (2017).

55. Ethayarajh, K. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *arXiv [cs.CL]* (2019).

56. Heeger, D. J. Theory of cortical function. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1773–1782

(2017).

57. Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
58. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience* (2020) doi:10.1038/s41583-020-00395-8.
59. Hasson, U., Nastase, S. A. & Goldstein, A. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).
60. Hart, B. & Risley, T. R. Meaningful differences in the everyday experience of young American children. **268**, (1995).
61. Weisleder, A. & Fernald, A. Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychol. Sci.* **24**, 2143–2152 (2013).
62. Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M. & Lyons, T. Early vocabulary growth: Relation to language input and gender. *Dev. Psychol.* **27**, 236–248 (1991).
63. Tan, H. & Bansal, M. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. *arXiv* (2020).
64. Chomsky, N. ASPECTS OF THE THEORY OF SYNTAX. (1964) doi:10.21236/ad0616323.
65. Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L. & Lewis, M. Generalization through Memorization: Nearest Neighbor Language Models. *arXiv [cs.CL]* (2019).
66. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16**, 199–231 (2001).
67. Bybee, J. & McClelland, J. L. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review* vol. 22 (2005).
68. The ‘Five Graces Group’ *et al.* Language Is a Complex Adaptive System: Position Paper. *Language Learning* vol. 59 1–26 (2009).
69. Goldberg, A. E. *Explain Me This: Creativity, Competition, and the Partial Productivity of*

Constructions. (Princeton University Press, 2019).

70. Hasson, U., Egidi, G., Marelli, M. & Willems, R. M. Grounding the neurobiology of language in first principles: the necessity of non-language-centric explanations for language comprehension. *Cognition* **180**, 135–157 (2018).
71. Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci. U. S. A.* (2020) doi:10.1073/pnas.1907367117.
72. Clark, K., Khandelwal, U., Levy, O. & Manning, C. D. What Does BERT Look at? An Analysis of BERT's Attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (2019) doi:10.18653/v1/w19-4828.
73. Mamou, J. *et al.* Emergence of Separable Manifolds in Deep Language Representations. *arXiv [cs.CL]* (2020).
74. Marcus, G. F. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. (MIT Press, 2019).
75. Yuan, J. & Liberman, M. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* **123**, 3878 (2008).
76. Yang, A. I. *et al.* Localization of dense intracranial electrode arrays using magnetic resonance imaging. *NeuroImage* vol. 63 157–165 (2012).
77. Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J.-M. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011**, 156869 (2011).
78. Lachaux, J. P., Rudrauf, D. & Kahane, P. Intracranial EEG and human brain mapping. *J. Physiol. Paris* **97**, 613–628 (2003).
79. Michelmann, S. *et al.* Data-driven re-referencing of intracranial EEG based on independent component analysis (ICA). *J. Neurosci. Methods* **307**, 125–137 (2018).
80. Jia, X., Tanabe, S. & Kohn, A. Gamma and the Coordination of Spiking Activity in Early

Visual Cortex. *Neuron* **77**, 762–774 (2013).

81. Manning, J. R., Jacobs, J., Fried, I. & Kahana, M. J. Broadband shifts in local field potential power spectra are correlated with single-neuron spiking in humans. *J. Neurosci.* **29**, 13613–13620 (2009).
82. Honey, C. J. *et al.* Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423–434 (2012).
83. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
84. Hall, P. & Wilson, S. R. Two Guidelines for Bootstrap Hypothesis Testing. *Biometrics* vol. 47 757 (1991).
85. Golovin, D. *et al.* Google Vizier. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017) doi:10.1145/3097983.3098043.