Thinking ahead: spontaneous next word predictions in context as a keystone of language in humans and machines

Ariel Goldstein[1,2Ŧ], 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], Sasha Devore[3], Adeen Flinker[3], 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

* Equal contribution

[Ŧ]Corresponding author: ariel.y.goldstein@gmail.com

Abstract

Departing from rule-based linguistic models, advances in deep learning resulted in a new type of autoregressive deep language models (DLMs). These models are trained using a self-supervised next word prediction task. We provide empirical evidence for the connection between autoregressive DLMs and the human language faculty using spoken narrative and electrocorticographic recordings. Behaviorally, we demonstrate that humans have a remarkable capacity for word prediction in natural contexts, and that, given a sufficient context window, DLMs can attain human-level prediction performance. Leveraging on DLM embeddings we demonstrate that the brain constantly and spontaneously predicts the identity of the next word in natural speech, hundreds of milliseconds before they are perceived. Finally, we demonstrate that contextual embeddings derived from autoregressive DLMs capture neural representations of the unique, context-specific meaning of words in the narrative. Our findings suggest that DLMs provides a novel biologically feasible computational framework for studying the neural basis of language.

Introduction

Modeling the underlying neural basis for language is central for understanding human cognition. Modeling language, however, is particularly challenging given that every language contains regularities, subregularities, and exceptions, which are conditioned by discourse context, meaning, dialect, genre, and many other factors. For example, English nouns are commonly pluralized by adding a final /-s/, but such a rule is violated across many contexts. For instance, types of fish (*salmon*) commonly remain unchanged in the plural, names of lower body-wear (*pants*, *leggings*) are grammatically plural even though they are semantically singular, and *rice* and *hair* are conventionally treated as mass nouns and so require no plural (*eat rice, pull hair*). There are exceptions to the exceptions as well. For instance, in certain contexts, *hair* can be treated as a singular or plural rather than mass (*a gray hair/gray hairs*). Against this rich backdrop, classical language models struggle to find a set of simple, yet generalizable linguistic rules that can be implemented across all contexts[1,2]. Usage- (context-) based constructionist approaches to language take this complexity into account[3–5]. By adopting the constructionist perspective, recent advances in deep learning led to the development of an entirely new family of deep language models (DLMs[6–15]). These models aim to bypass the need to learn concise, fully generalizable, rules by learning to predict appropriate usage based on how speakers have used language in similar past contexts[16–18].

From a linguist's perspective the applied success of DLMs is striking because they rely on a very different architecture than classical language models[19]. Traditional investigation of the neural basis of language relied on classical language models. These models employ symbolic linguistic elements (like nouns, verbs, adjectives, adverbs, determiners), and combine them with rule-based operations embedded in hierarchical tree structures[20–22]. In contrast, DLMs do not parse words into parts of speech, but encode words using continuous numerical vectors (arrays of real numbers termed **embeddings**) that can be combined using a series of simple arithmetic operations (e.g., sequential matrix multiplication as opposed to complex syntactic rules). These word embeddings are learned from real-world textual examples "in the wild," with minimal prior knowledge about the structure of language. Curically, learning is guided by a simple self-supervised autoregressive objective, with the sole aim to predict the next word in the context of preceding words[6,9,11,12,14]. To the surprise of many, the next word prediction task, by itself, was sufficient to generate well-formed linguistic outputs[13,23,19].

The engineering feat of DLMs has transformed the way people interface with computers. However, it is unclear whether these prediction-based self-supervised DLMs relate to the way the human brain processes language. By mapping between language and the corresponding neural activity, recent studies have harnessed machine learning to decode linguistic information from the brain[24–32]. Taking a step forward, some theoretical and empirical work, especially in visual neuroscience[18,33–35], but recently also in language[25–31,36], posits that deep neural networks may provide a new modeling framework to study neural computations in

biological neural networks. To substantiate the call for this paradigm shift, we provide new behavioral and neural evidence for the connection between prediction-based DLMs and the human brain as they process natural speech.

Autoregressive DLMs incorporate two key principles: 1) they learn in a self-supervised way by generating next-word predictions; 2) they rely on a large trailing window of context to generate their predictions. Here we explore the hypothesis that human language in natural settings also abides by these two fundamental principles of *prediction* and *context*. To test the first hypothesis, we tested whether human listeners can predict each word in natural contexts in a way that matches DLM prediction; and that the brain spontaneously predicts each word before it is articulated. So far, the ability of listeners' to predict upcoming words has been tested with highly-controlled sentence stimuli (i.e., cloze procedure)[37–41]. For that end, we developed a new behavioral paradigm to test listeners' ability to predict upcoming words as they listen to a 30-minute spoken story. We demonstrate that humans have a remarkable capacity for word prediction in natural contexts, and that, given a sufficient context window, DLMs can attain human-level prediction performance. Neurally, we demonstrate that the human brain spontaneously generates next-word predictions as it listens to the same 30-minute story. Many studies have reported neural signals associated with word level prediction **after** word onset using controlled[42–46] as well as naturalistic stimuli[28–30,47]. Leveraging high-precision electrocorticographic (ECoG) recordings and DLM embeddings, we demonstrate that the brain spontaneously predicts the meaning of upcoming words hundreds of milliseconds before these words are perceived. Furthermore, analysis of instances where the predicted words did not match the identity of the perceived words allowed us to dissociate pre word onset neural processes associated with next word prediction from post word onset neural processes of the incoming perceived words. Together, these results clearly separate pre-word onset processes associated with word prediction from post-word onset neural processes. Finally, we tested the second hypothesis, that the brain, similarly to DLMs, encods the trailing context by relying on contextual embeddings. Indeed, contextual embeddings derived from autoregressive DLMs improved our ability to model neural responses to words in natural spoken language above static embeddings. Together, our findings provide compelling evidence for the deep connections between artificial neural network models and the human brain and support a new modeling framework for studying the neural basis of the human language faculty.

Results

Word-by-word behavioral prediction during a natural story

DLMs can be effectively trained using simple, self-supervised objectives, such as next-word prediction. Using this objective, the model learns to predict each upcoming token based on the preceding context in a text (also known as autoregressive DLMs). This self-supervised learning objective may be relevant to human learning, as feedback about the accuracy of next word prediction is readily available to all listeners from infancy to adulthood as they listen to natural speech. Listeners' ability to predict forthcoming words in natural speech, however, is largely unknown. Behavioral studies have mainly focused on listeners' ability to predict the last word in highly-controlled sets of isolated sentences or used indirect measures such as eye-tracking and reaction time[37-48].

To address this gap, we developed a novel behavioral paradigm, which enabled us to assess the human ability to predict each upcoming word in a natural context. To obtain a continuous measure of people's ability to predict the next word in the narrative, we used a sliding-window behavioral experiment. In this experiment, 50 participants attempted to predict every upcoming word of a 30-minute podcast (see Methods and Materials), "Monkey in the Middle" by This American Life[48] (Fig. 1A-B). This procedure provides 50 predictions for each of the story's ~5000 words (see Fig. 1C, and Materials and Methods section for further details). Next, we calculated a mean prediction performance (proportion of participants predicting the correct word) for each word in the narrative, which we refer to as "predictability score" (Fig. 1D). A predictability score of 100% indicates that all subjects correctly guessed the next word and predictability score of 0% indicates that no participant predicted the upcoming word.

Overall, participants were able predict many upcoming words in a complex and unfamiliar story (mean predictability score across all words = 28%, SE = 0.5%). The predictability score for blindly guessing the most frequent word in the text ("the") is 6%. Human capacity for prediction was even more impressive when focusing on the predictability of individual words (Fig. 1D). About 600 words had a predictability score higher than 70%. Interestingly, high predictability was not confined to the last words in a sentence, and included words from all parts of speech (21.44% nouns, 14.64% verbs, 2.24% adjectives, 41.62% functions words, 2.11% adverbs, 17.94% other).

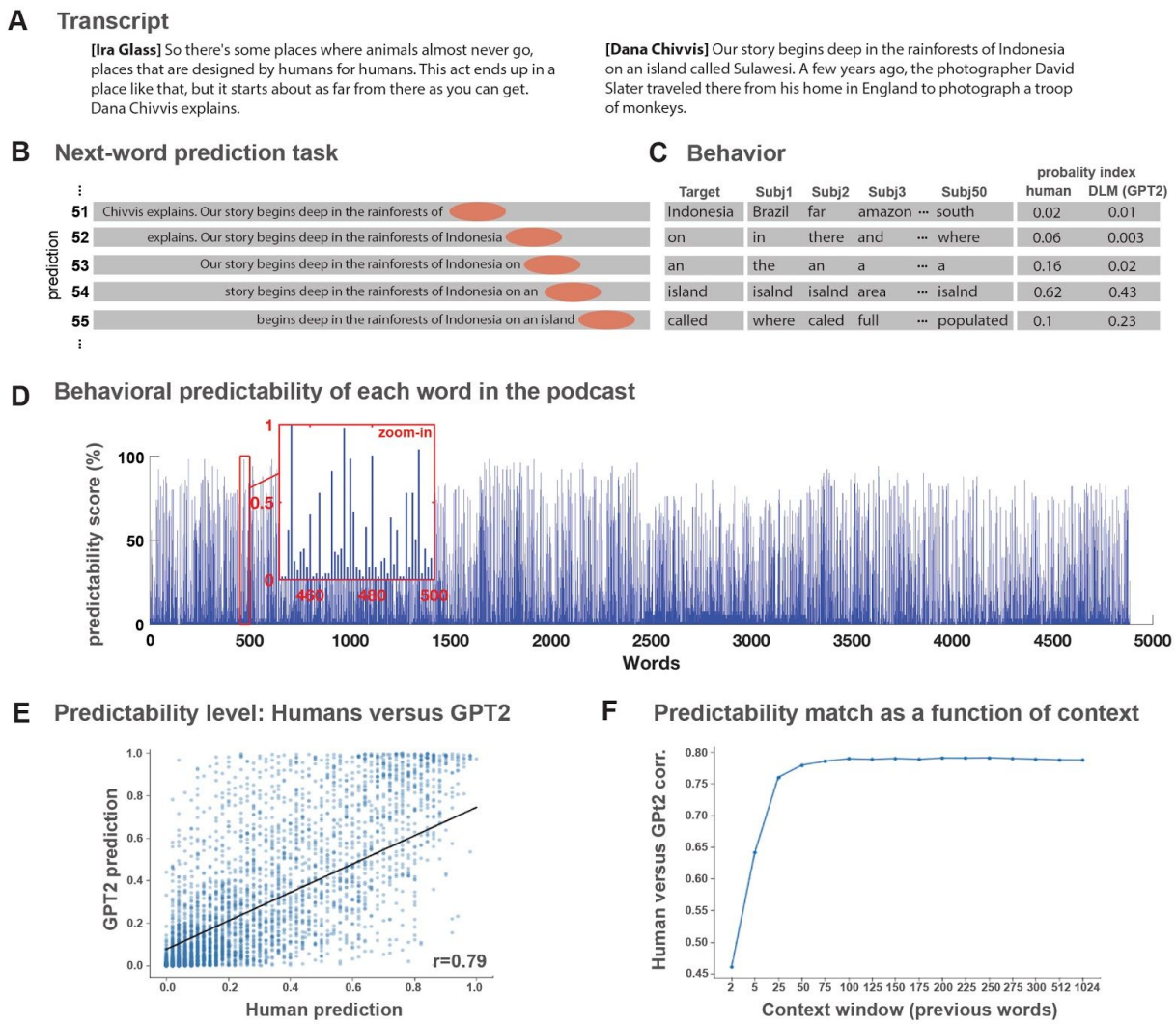


Figure 1. Behavioral assessment of the human ability to predict forthcoming words in a natural context. A) The stimulus was transcribed for the behavioral experiment. **B)** A 10-word sliding window was presented in each trial and participants were asked to type their prediction for the next word. Once typed, the correct word was presented, the window was moved forward by one word, and participants were asked to predict the subsequent word. **C)** For each target word, we calculated the proportion of participants that predicted the forthcoming word correctly. **D)** Predictability scores varied considerably across words. **E)** For each word in the story, we compared the human predictability score (*x*-axis) and the probability GPT2 assigned to the correct word (*y*-axis). **F)** We calculated the correlation between human predictions and GPT2 predictions (as reported in panel D) when supplying GPT2 with different context windows ranging from 2–1024 preceding words.

Next, we compared human performance in predicting the next word in natural speech to DLM performance (Fig. 1E-F). We used GPT2, an autoregressive DLM trained to predict the next word in large corpora of real-world text, to obtain the predictability for each word in the story as a function of its context. 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 _____." Remarkably, human and GPT2 estimates of predictability were highly correlated (Fig. 1E, $r = .79$, $p < .001$).

Timescale of word prediction

In natural comprehension (e.g., listening to a story) predictions for upcoming words may be influenced by information accumulated over multiple timescales. Time scales that range from the most recent words to information gathered over multiple paragraphs as a story unfolds [49]. Next we varied GPT2's input window size (from two words up to 1024 words) 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).

Neural markers of word prediction

Given the participants' remarkable capacity for next-word prediction, in the second study we searched for evidence of next-word prediction before word onset at the neural level. Electrocorticography (ECoG) data with high spatial and temporal resolution was recorded from eight epileptic patients, who volunteered to participate in the study (see Fig. 2A for a map of all electrodes). We had better coverage in the left hemisphere (917 electrodes) than in the right hemisphere (233 electrodes). Thus, in the paper body we mainly focus on language processing in the left hemisphere, but for exhaustiveness we also present maps for the right hemisphere in supplementary materials. All participants listened to the same spoken story used in the behavioral experiment. The participants engaged in free listening with no explicit instructions to predict upcoming words. Comprehension was verified using a post listening questionnaire.

To model the neural responses to each word, before and after word onset we used a linear encoding model to learn the mapping between an embedding space and the neural responses (see Fig. 2B, and Materials and Methods section for details). There are two types of word embeddings: *static embeddings* (e.g., word2vec[50], GloVe[51] and others[52]), which assign a single vector to each word in the lexicon (e.g., "*cold*") irrespective of context. And *contextual embeddings* (BERT[8] , GPT-1,2,3[12,13] and others[12,13,6,7,9,53,54]) in which the same word is assigned different embeddings (vectors) as a function of the surrounding words (e.g., "*cold*" receives different embeddings in the context of "it is freezing *cold*" versus "you are nasty and *cold*"). The results from each encoding model were thresholded with correction for multiple tests using false-discovery rate (FDR) at $p < .01$ (Materials and Methods section).

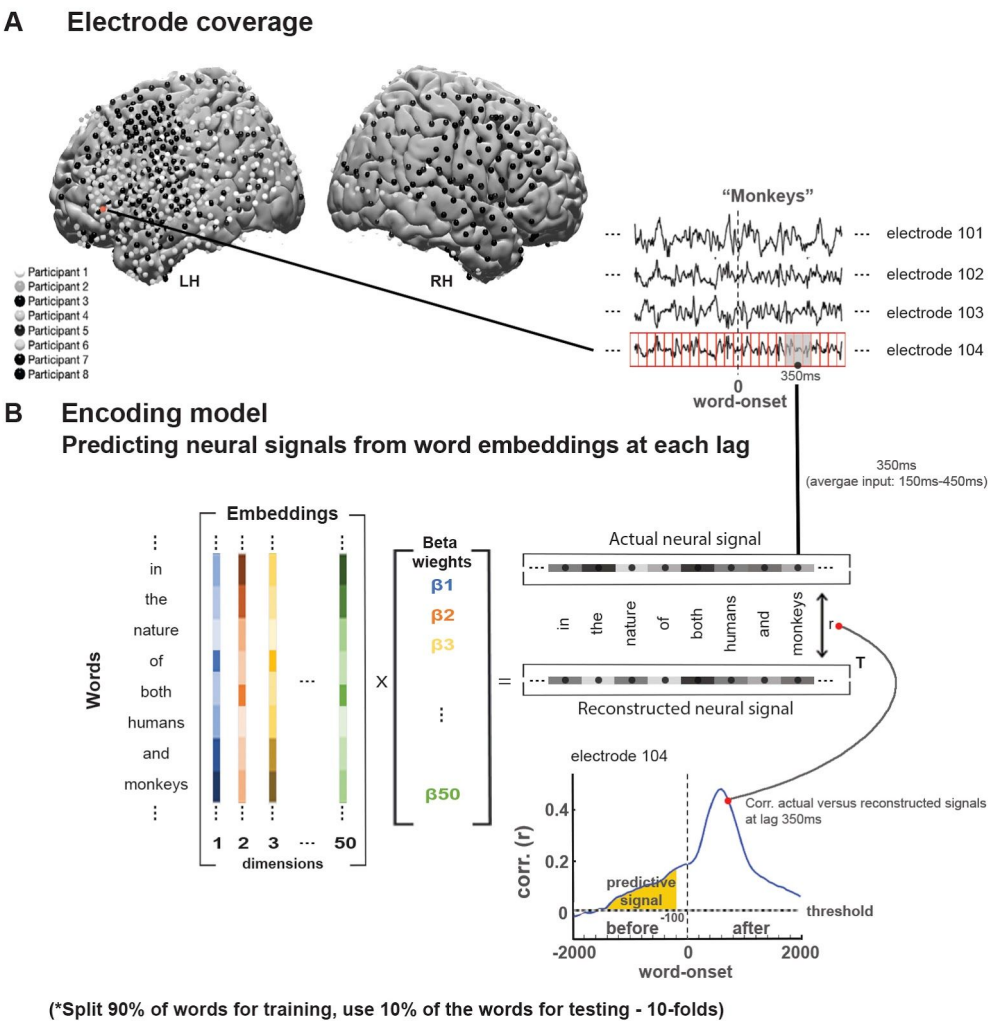


Figure 2. Linear encoding model used to predict the neural responses to each word in the narrative before and after word onset. A) Brain coverage consisted of 1086 electrodes from eight participants. The words were aligned with the neural signal such that the onset of each word (moment of articulation) is at lag 0. Responses for each word were averaged over a window of 200 ms and provided as input to the encoding model. **B)** In the encoding model, each word is represented by a 50-dimensional vector (embedding). In most analyses, we used three types of embeddings: arbitrary embeddings (random vectors), static embeddings (from GloVe), and contextual embeddings (from GPT2). For each temporal shift (lag), all words were split into non-overlapping training and test folds for cross-validation. In a single cross-validation fold, 90% of the instances of words were used to train the encoding model and 10% of the words were used to test the encoding model. A series of 50 coefficients corresponding to the features of the word embeddings was 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 for 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).

Using arbitrary embeddings to predict neural responses to natural speech
First, we trained encoding models to predict neural responses using 50-dimensional arbitrary embeddings (vectors). Arbitrary embeddings (sampled from uniformed [-1,1] distribution) provide a minimalist model to search for electrodes in which each word has a consistent, separable, neural response profile. These embeddings do not encode any information about the semantic or syntactic relations between words in the lexicon, or about the contextual dependencies among the specific sequence of words in the podcast. The encoding analysis with arbitrary embeddings identified 57 electrodes in the left hemisphere (LH) with significant correlations (after correction for multiple comparisons). Electrodes were found in early auditory areas, motor cortex, and language areas (see Fig. 3A for LH electrodes, and supplementary Fig. S1A for right hemisphere (RH) electrodes). The arbitrary embeddings provide us with a baseline measure of word-level encoding performance for a model that is deprived of any information about statistical relations among words.

Using static word embeddings to predict neural responses to natural speech
Using *static* word embeddings (GloVe-50d) substantially improved the ability of the encoding model to predict the neural responses for each word (Fig. 3B, S1B). Models like GloVe and word2vec learn a single, static, embedding for each word based on the co-occurrence of neighboring words over a large training corpus. These static embeddings capture some semantic and syntactic properties of language: for example, they capture semantic relationships like "*Paris* is to *France* as *Rome* is to *Italy*" and grammatical relationships like "*walked* is to *walk* and *ran* is to *run*"[50]. Compared to the arbitrary embeddings, using static embeddings informs us whether the compressed statistical structure of real-world language better captures neural responses to words. GloVe substantially improved encoding model performance in predicting the neural responses to unseen words. GloVe based encoding resulted in statistically significant correlations in 129 electrodes in LH (Fig. 3B, and 25 electrodes in RH, Fig. S1). 88 of these electrodes (75 in LH) were not captured with the arbitrary embeddings. The additional electrodes were located in the inferior frontal cortex, primary and supplementary motor cortex, and temporal cortex extending from the angular gyrus to the temporal pole, including early auditory cortex and lateral auditory association cortex. Peak model performance improved from .2 for arbitrary embeddings to .49 for static embeddings.

Using contextual embeddings to predict neural responses to natural speech
Replacing static (GloVe) with contextual embeddings (GPT2) further improved the ability of the encoding model to predict the neural responses to words (Fig. 3C, S1C). Autoregressive DLMs, such as GPT2 [12], learn contextual word embeddings by relying on a simple objective function of minimizing next-word prediction error in text corpuses. As opposed to models that learn static word embeddings (e.g., GloVe), GPT2 is trained to assign a unique vector to each token based on preceding words in the story (up to a contextual window of 1024 tokens). In other words, GPT2 is trained to compress the information of prior words and output a contextual

embedding for assessing the next word probabilities. Thus, next we asked whether GPT2-based contextual embeddings provide a further improvement in model performance over GloVe. Before fitting the encoding model with GPT2, we used PCA to reduce the dimensionality of GPT2 to match the 50-dimensionality of GloVe embeddings. Using contextual embeddings (GPT2) substantially improved the ability of the encoding model to predict the neural responses (Fig. 3C). Encoding based on contextual embeddings (GPT2) resulted in statistically significant correlations in 164 electrodes in LH (and 34 in RH). 71 of these electrodes (57 in LH) were not captured with the static embeddings (GloVe). The map in Fig. 4A summarizes the results. Electrodes with robust encoding only for GPT2 embeddings are marked in purple. Electrodes with robust encoding for GloVe and GPT2 embeddings are marked in yellow. and electrodes with significant encoding for all three types of embeddings are marked in red (Fig. 4A). The additional electrodes revealed by the contextual embedding (purple) were located in the inferior frontal gyrus, primary and supplementary motor cortices, temporal pole, posterior superior temporal gyrus, parietal lobule, and angular gyrus (Fig. 4A).

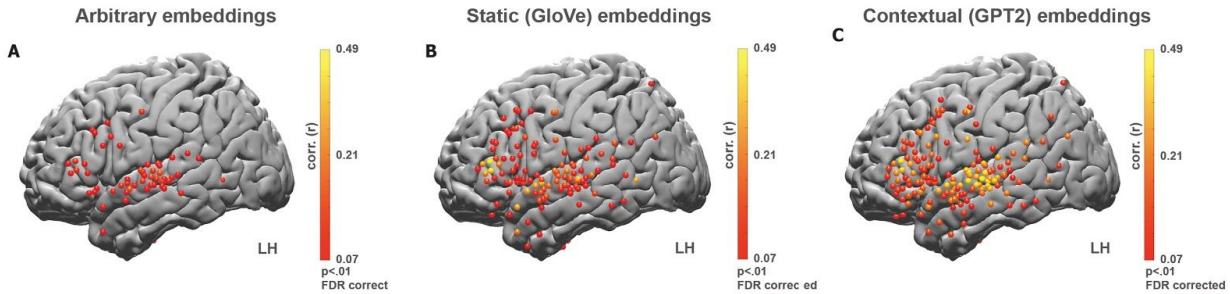


Figure 3. Contextual (GPT2) embeddings outperform static (GloVe) and arbitrary embeddings in predicting neural responses to words in the narrative. A) Peak correlation across lags between predicted and actual word responses for the arbitrary embeddings (nonparametric permutation test; $p < .01$, FDR corrected). **B)** Peak correlation between predicted and actual word responses for the static (GloVe) embeddings. **C)** Peak correlation between predicted and actual word responses for the contextual (GPT2) embeddings. Note that using contextual embeddings significantly improved the ability of the encoding model to predict the neural signals for unseen words across many electrodes.

Encoding neural responses before and after word onset

In the behavioral experiment (Fig. 1) we observed people's remarkable capacity to predict upcoming words in the story. Next, we tested whether the neural signals also contain information about words before and after they are perceived (i.e., word onset). To that end, we re-estimated the encoding model at different lags between -5000 ms and +5000 ms relative to word onset using a sliding window at increments of 25 ms (see Fig. 2 as well as Methods and Materials). A sample of single-electrode encoding results across all lags is presented in Fig. 4A, and the average model performance across different subsets of electrodes is presented in Fig. 4B-D.

We obtained better encoding performance for static embedding (GloVe) than for arbitrary embeddings in numerous electrodes (for example at IFG, PostCG, AG and TP) up to 1 sec before word onset (Fig. 4A). The pattern is seen in the average responses across electrodes, with significant encoding for all three types of embeddings (GPT2, GloVe and arbitrary, Fig. 4B), as well as in the average model performance across electrodes with significant encoding for both Glove and GPT2 (but not arbitrary) embeddings (Fig. 4C). The peak of encoding for both arbitrary and static embeddings was observed 150–200 ms after word onset (lag 0), but the models performed above chance up to -1000 ms before and 1000 ms after word articulation (as indicated by the red horizontal threshold). This supports the claim that the brain has predictive information about the upcoming word before it is perceived . Furthemore, the encoding performance for the static embeddings (GloVe) was statistically higher than for the arbitrary embeddings in all electrodes (Fig. 4C). Crucially, the improvement in prediction performance for the static embedding model over the arbitrary embeddings (Fig. 3, 4) cannot be attributed to wholesale geometrical properties of the GloVe vector space, as the improvement was abolished when we shuffled the assignment of the GloVe embeddings to words, thus removing the relational linguistic information from the model (Fig. S2). Furthermore, the performance before word onset cannot be attributed to correlations between adjacent word embeddings in the story or to the existence of bigrams, as it was not affected by the removal of the previous GloVe embedding from the current GloVe embedding before running the encoding analysis (Fig. S3). The encoding results using GloVe embeddings were replicated using 100-dimensional static embeddings from word2vec (Fig. S4). Together, these results unequivocally demonstrate that neural responses to words in the spoken story, before and after word onset, are better modeled by an embedding space that captures the statistical relations between words in natural language.

Contextual embeddings (GPT2) provided an additional improvement in the encoding performance. The improvement is seen both at the peak of the encoding and the width of the encoding (Fig. 4B-D), which can reach up to four seconds before words-onset in the most selective electrodes (Fig. 4C-D). Such results can clearly be seen both in single electrodes (Fig. 4A) and in the aggregate across electrodes' averages (Fig. 4B-D). Together, these results demonstrate that neural responses to words in the spoken story are better modeled by contextual embeddings, which capture the unique contextual meaning of each word. The improvement in the encoding model for the contextual embeddings (GPT2) was robust and apparent across the three subset selections of electrodes (Fig. 4B-D).

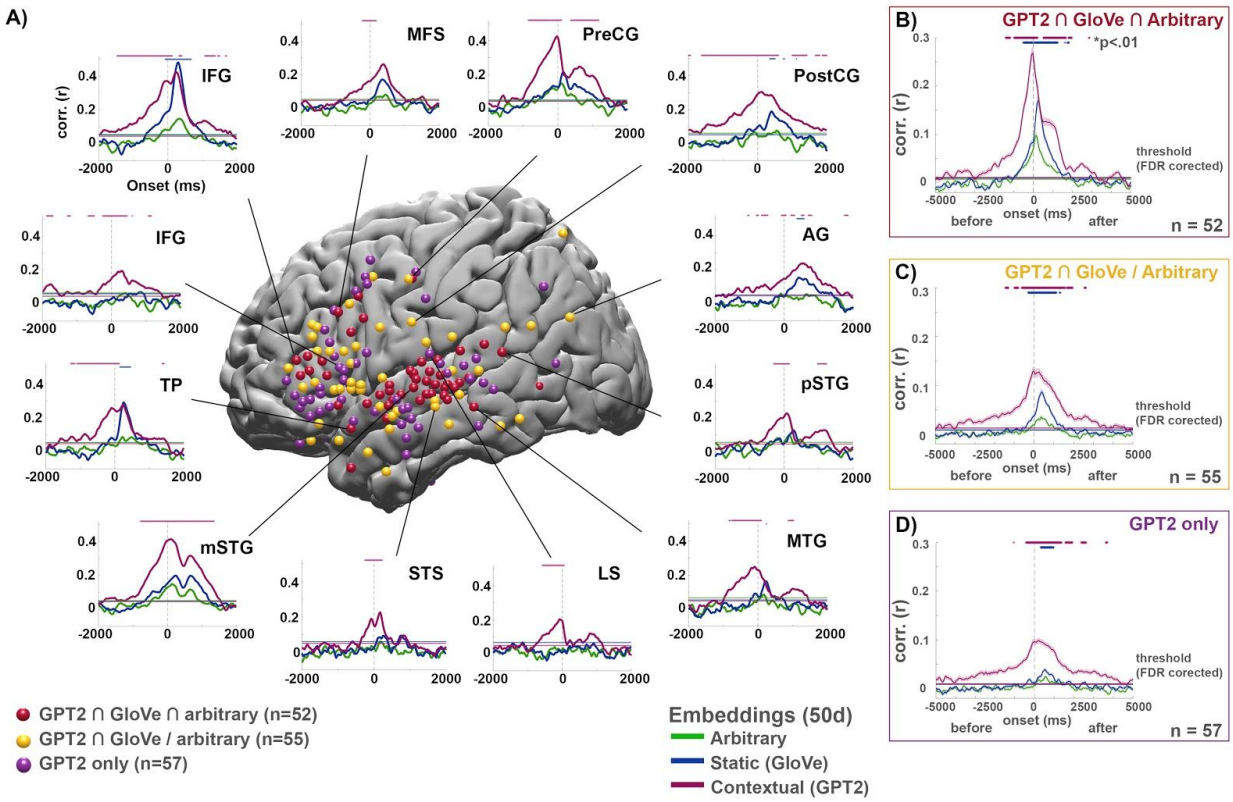


Figure 4. Contextual embedding significantly improves the modeling of the neural signals. A) Map of the electrodes in the left hemisphere with significant encoding for: 1) all three types of embeddings (GPT2 ∩ GloVe ∩ arbitrary, red); 2) for static and contextual embeddings (GPT2 ∩ GloVe, but not arbitrary, yellow); 3) and contextual only (GPT2, purple) embeddings. Note the three groups do not overlap. 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 electoreds 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 electoreds marked in yellow). **D)** Average encoding model performance across lags for all electrodes with significant encoding for GPT2-only (57 electoreds 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).

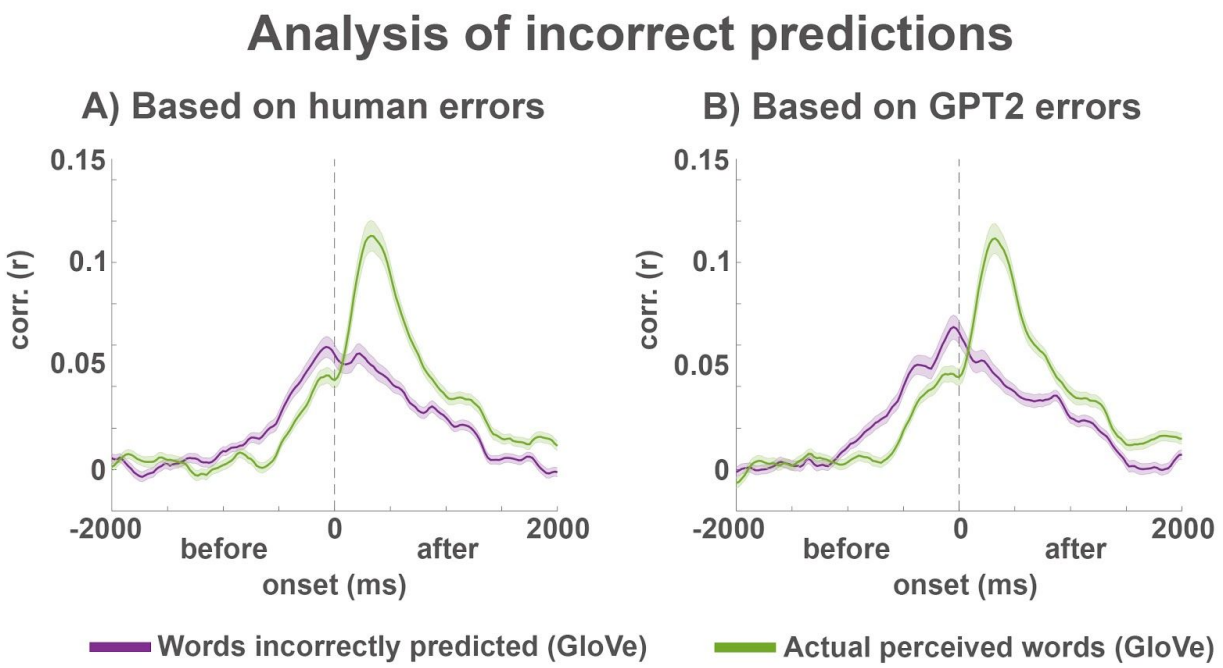


Figure 5. Analysis of incorrect predictions dissociates neural processes associated with next-word prediction from post-word onset processes of incoming words. A) Encoding model (using static GloVe embeddings) for incorrect top-1 human predictions (purple), and for the actual words listeners perceived in these exact same instances (green). **B)** Same analysis as in A, using GPT2's top-1 incorrect predictions. Note that using the embeddings of the predicted words improved our ability to model the neural responses prior to word onset, while using the embeddings of the actual perceived words improved the ability to model neural responses after word onset.

Next word prediction versus processing of incoming words

To disentangle neural processes associated with next-word prediction from neural processes of incoming words, we next focused on incorrect predictions. To that end we analyzed instances where humans' top-1 predictions before word onset (Fig 1D) did not match the actual perceived word after word onset. Top-1 predictions were defined by the most frequently predicted word across all participants. Here we only analyzed those instances where the most predicted word was incorrect. Focusing on the incorrect predictions allowed us to model the same neural signals before and after word onset, using two GloVe embeddings of: 1) what subjects incorrectly predicted; 2) what subjects actually perceived.

Using the word embedding of the incorrect predictions improved our ability to model the neural responses before word onset, with maximal encoding just before lag 0 (Fig. 5A, purple). This demonstrates that the pre-onset neural activity contains information about human expectations, irrespective of what listeners actually perceived. Similar results were obtained when we used GPT2's incorrect top-1 predictions (Fig. 5B). This demonstrates that GPT2

predictions provide an adequate cognitive model for people's spontaneous and implicit predictions. In contrast, using the word embeddings of the actual words listeners perceived, improved our ability to model the neural responses after word onset (with maximal encoding 200-300ms after lag 0, Fig. 5A-B, green). These results disentangle pre-word onset processes associated with word prediction from post-word onset processes, and further support the claim that GPT2 and human next-word predictions mirror each other.

Modeling the context versus predicting the upcoming word

The improved prediction of neural responses before word onset using GPT2 can be attributed to two related factors that are absent in the static (GloVe-based) word embeddings: 1) GPT2 actively predicts the upcoming word in the story; and 2) GPT2 aggregate information about the preceding words in the story which yield unique, context-specific, embeddings. By carefully manipulating the contextual embeddings and developing an embedding-based decoder, we show how both context and next-word prediction contribute to the improved ability of GPT2 over GloVe to model the neural responses.

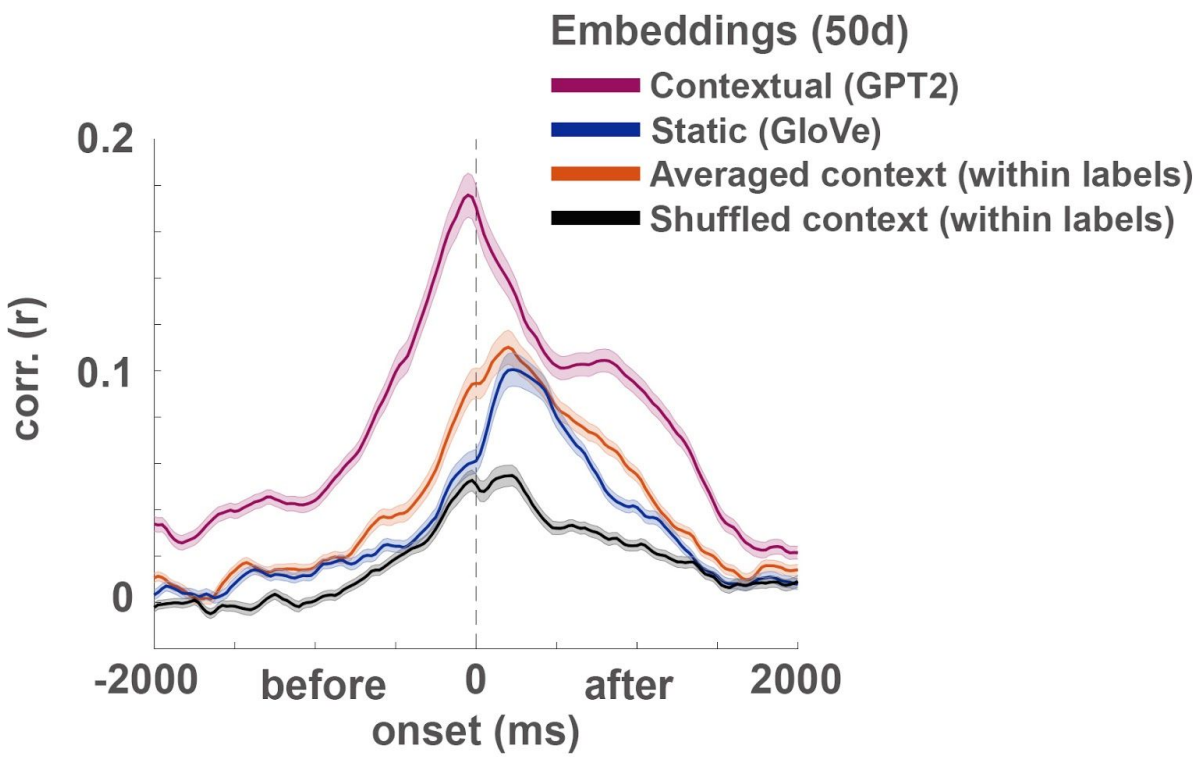


Figure 6. Contextual (GPT2) embeddings further improve the modeling of neural responses prior to word onset. Encoding model performance for contextual embeddings (GPT2) aggregated across all electrodes with preferred encoding for GloVe (Fig. 3B): original contextual embeddings (purple), static embeddings (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., by concatenating) the GloVe embeddings for the prior words in each given 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 prior to word articulation (Fig. S5). This result suggests that GPT2's contextual embeddings are better suited to capture the contextual information embedded in the neural responses than GloVe.

A complementary way to test this hypothesis is to remove the unique contextual information from GPT2 embeddings. We removed contextual information from GPT2 embeddings by averaging the embeddings of all tokens of each of the unique words (e.g., all occurrences of the word "monkey") into a single vector. Thus, we collapsed the contextual embedding into a static embedding (similar to GloVe) in which each unique word in the story is represented by one unique vector. Note that the resulting embeddings are still specific to the overall context of this particular podcast (unlike GloVe), but 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 GPT2 by averaging occurrences of each given word effectively reduced the performance of the encoding model to that of the static GloVe embeddings (Fig. 6, orange).

Finally, we examined the specificity of the contextual information contained in both GPT2 embeddings and neural responses by scrambling the embeddings for each word across different occurrences of the same word in the story (e.g., switch 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 occurrences of the same word across contexts substantially reduced encoding model performance (Fig. 6, black), pointing at the high contextual dependency represented in the neural signals. Taken together, these results suggest that contextual embeddings provide us with a new way to model neural representation of the unique, context-dependent, meaning of words occurring in natural contexts.

Using autoregressive DLMs to better predict the next word before articulation

To further test whether contextual embeddings improve our ability to model the predictive neural responses prior to word onset, we turned to a decoding analysis. The encoding model finds a mapping from the embedding space to the neural responses in an attempt to predict

neural responses to novel words, not seen in training. The decoding analysis inverts this procedure to find a mapping from neural responses to the embedding space with the goal of predicting the identity of words in new contexts by aggregating data across space (electrodes) and time[55]. This analysis adds a crucial layer to the encoding model. It allows us to directly quantify the amount of aggregate neural information, before or after word onset, as to the meaning of words we can extract when relying on each type of embedding. The decoding analysis was performed in two steps. First, we trained a deep convolutional neural network to aggregate neural responses within a 625 ms window (Fig. 7A) and mapped this neural signal to the arbitrary, GloVe, and GPT2 embedding spaces (Fig. 7B). To not confer an advantage to GPT2 over GloVe, we used a set of electrodes with significant encoding for both types of embeddings as input to the decoding model (the union of 107 electrodes in Fig. 4B-C). Second, the predicted word embeddings were used for word classification based on their cosine-distance from all embeddings in the dataset (Fig. 7C). Although we evaluated the decoding model using classification, the classifier predictions were constrained to rely only on information contained within the embedding space. This is more conservative than an end-to-end word classification approach, which may, for example, capitalize on acoustic information in the neural signal that is not encoded in the language models.

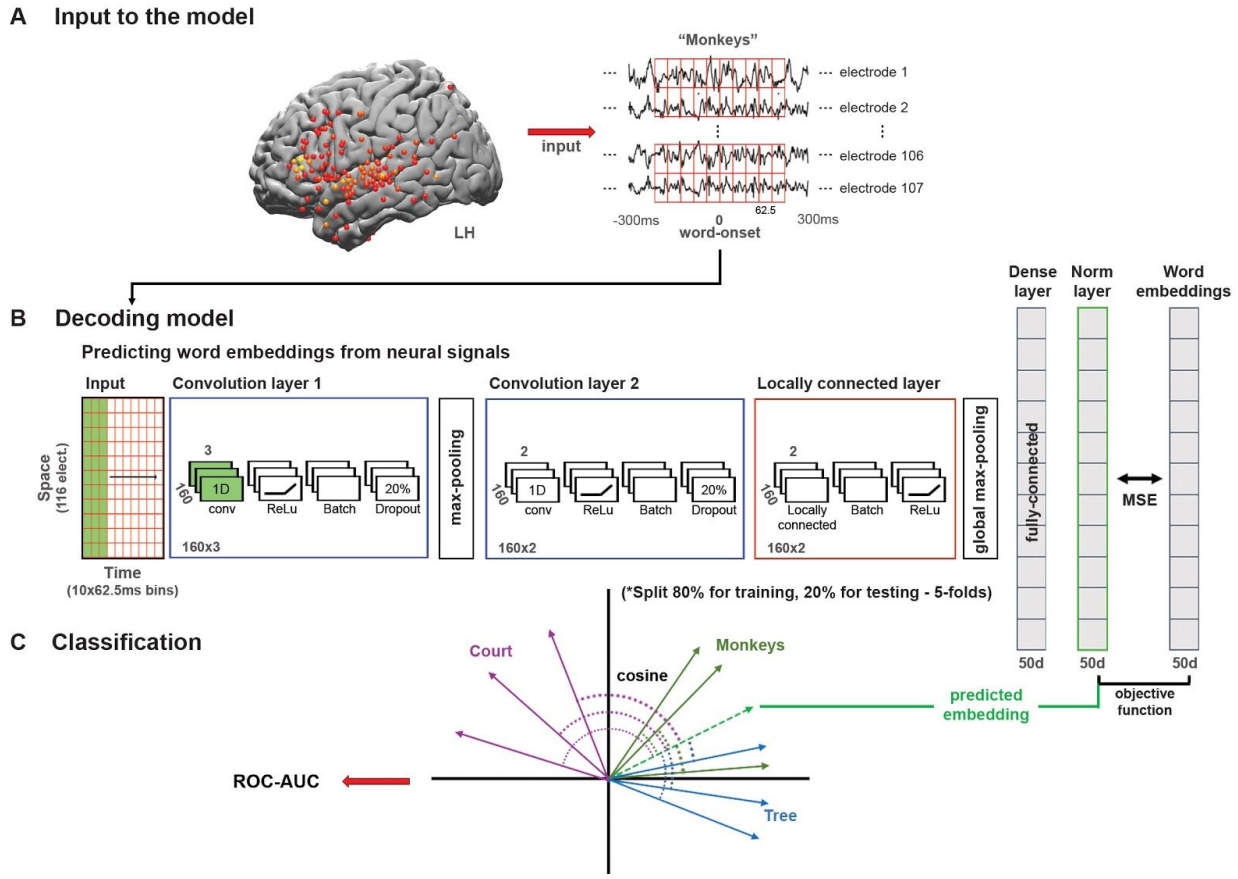


Figure 7. Deep nonlinear decoding model used to predict words from neural responses before and

after word onset. A) Neural data from left hemisphere electrodes with significant encoding model performance using GloVe embeddings (from Fig. 4BC) were used as input to the decoding model. The stimulus is segmented into individual words and aligned to the brain signal at each lag. The signal is then averaged within 62.5-ms bins spanning a temporal window of 625 ms (10 62.5 ms bins). **B)** A nonlinear decoding model was trained to predict the word embeddings from the neural signals. 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 based on the embedding space used to construct ROC-AUC scores. This enabled us to assess how much information about specific words can be extracted from the neural activity via the linguistic embedding space.

Using a contextual decoder greatly improved our ability to classify the identity of words over decoders relying on static and arbitrary embeddings (Fig. 8). 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 curve that falls on the diagonal line and suggests that the classifier does not discriminate between the words. In this case the AUC score of the classifier will be 0.5[56]. Classification using GPT2 outperformed GloVe and arbitrary embeddings both before and after word onset.

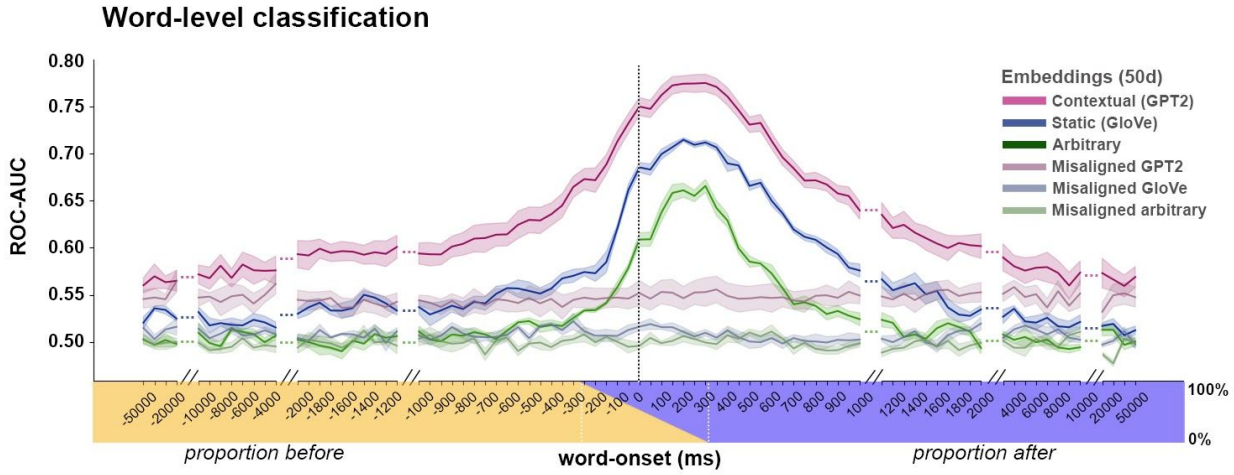


Figure 8. Using a decoding model for classification of words before and after word onset. *Word-level classification.* Classification performance for contextual embeddings (GPT2; purple), static embeddings (GloVe; blue), and arbitrary embeddings (green). The *x*-axis labels indicate the center of each 625 ms window used for decoding at each lag (between -50 to 50 sec). The colored strip indicates the proportion of pre- (yellow) and post- (blue) word onset time points contained in each lag. Dimed lines indicate decoding performance with misaligned (shifted) labels for the contextual (dim purple), static (dim blue), and arbitrary (dim green) embeddings. Error bands denote SE across five test folds. Note that contextual embeddings improve classification performance over GloVe both before and after word

onset.

A closer inspection of the GPT2-based decoder indicates that the classifier managed to detect reliable information as to the identity of words several hundred milliseconds before word onset (Fig. 8). In particular, starting at about -2000 ms prior to word onset, when the neural signals were integrated across a window of 625 ms, the classifier detected predictive information as to the identity of the next word. The information about the identity of the next word gradually increased and peaked at an average AUC of 0.77 at a lag of 150 ms after word onset, when the signal was integrated across a window from -150 ms to 450 ms. Moving further beyond 300 ms from word onset, information about the exact word identity gradually declined. GloVe embeddings show a similar trend with a marked reduction in classifier performance (Fig. 8, blue). The decoding model's capacity to classify words before word onset demonstrates that the neural signal contains a significant amount of predictive information about the meaning of the next word, up to a second before it is perceived. At longer timescales, up to a few seconds, GPT2 predictions were still above baseline, till it converged to a baseline level. Baseline level was computed by misaligning (temporally shifting) the GPT2 embeddings by 2500 words. The misaligned embeddings retain the slowly varying context, while removing the word-specific alignment with the neural signals. In this case, decoding performance was stable with an AUC of 0.55 across all lags (Fig. 8, light pink). Given that the context-blind GloVe embeddings tend to be less correlated with each other than GPT2 embeddings, time-shifted baseline performance for GloVe was consistently lower than GPT2 (AUC = 0.51), but still above 0.5. Arbitrary embeddings, by construction, do not retain any temporal relation across embeddings, setting the baseline for shuffled embeddings at AUC = 0.50. The qualitative advantage in classification for GPT2 embeddings still holds, even when it conservatively normalized for baseline (cross-correlation) differences among the three types of embeddings (Fig. S6).

Discussion

Deep language models (DLMs) provide a new modeling framework that drastically departs from classical language models (CLMs). 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 or other linguistic terms. Rather, they learn from surface-level linguistic behavior to predict and generate the contextually appropriate linguistic outputs. The results provide compelling behavioral and neural evidence for deep connections between autoregressive (predictive) language models and the human brain.

Next-word prediction

Autoregressive DLMs learn according to the simple self-supervised objective of context-based next word prediction. Before this study, the extent to which proficient English speakers would be able to predict each word in the context of continuous, minutes-long, natural speech was underspecified. Our behavioral results revealed a remarkable capacity for next-word prediction in real-life stimuli. Crucially, it matches the next word prediction capacity of modern autoregressive DLMs as GPT2 (Fig. 1). On the neural level, by carefully analyzing the temporally resolved ECoG responses to each word as subjects freely listened to an uninterrupted spoken story, our results suggest that the brain has the spontaneous propensity to predict the identity of upcoming words before they are perceived (Fig. 3-8). Predictive neural signals up to -1000 ms prior to word onset discovered in the current study expand on the characterization of post word onset neural signals associated with prediction error and surprise. Such delayed neural signals are usually found 200-600 ms after word onset (e.g., N400 and P600 event-related potentials) when expectations are violated in isolated sentences[42–46] , and natural contexts[28–30,47,57]. Furthemore, an analysis of the instances where humans, as well as DLMs, failed to predict the identity of the upcoming words, clearly demonstrated that the pre-word onset neural signals contain information about listeners' internal expectations. Finally, by relying on DLM's contextual embedding the decoding analysis demonstrates that the meaning of upcoming words can be read from the neural signals hundreds of milliseconds before they are perceived.

In classical psychology, next-word prediction is discussed in the context of efficient processing (e.g., priming and improved reaction time) and not in the context of learning[29,46,58,47]. In machine learning, in contrast, as implemented in the autoregressive DLMs, next-word prediction is primarily essential to guide learning. Such a learning procedure, however, relies on gradually exposing the model to millions of real-life examples. Our finding of spontaneous predictive neural signals as subjects listen to natural speech suggests that next-word prediction may support lifelong learning. Furthemore, observational work in developmental psychology suggests that each day children are exposed to tens of thousands of words in contextualized speech each day, making the volume data available for learning quite large[59–61]. Future studies, however, should assess whether these cognitively plausible, prediction-based, feedback signals are indeed available at a young age as we learn language, and whether the brain can

use such predictive signals to guide language acquisition. Further, while next-word prediction may be an effective learning objective for language acquisition, it is not the only feasible objective—the brain may optimize additional simple objectives, at different timescales, to facilitate learning[18,62].

Contextual embeddings

Language is fundamentally contextual, as each word attains its full meaning in the context of preceding words over multiple timescales[49]. 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[63,64]. The contextual word embeddings learned by DLMs provide a new way to compress linguistic context into a numeric vector space, based on language-use statistics. This ability to efficiently represent words and make predictions using the statistics of natural language is related to Shannon's classical work on compressing and prediction of information bandwidth based on the natural statistics of words in texts[65,66]. Using a numerical vector space to represent neural responses to linguistic information avoids the circularity built in many classical language models that rely on linguistic terms to explain how language is encoded in neural substrates[67,68]. The results indicate that compressing the semantic and syntactic information into static embeddings (e.g., GloVe) greatly improves the ability to predict neural responses in many language areas[57] along the superior temporal cortex, parietal lobule and inferior frontal gyrus. Furthermore, the switch from static to contextual embeddings boosted our ability to model neural responses during the processing of natural speech across many brain areas. Taken together, these results suggest that the contextual embeddings learned by autoregressive DLMs provide valuable information as to the way the brain codes words in context.

Circuit architecture

We hypothesize that the family of DLMs shares certain critical features with biological language. This does not imply that they are identical, nor that they share the same circuit architecture. Human brains and DLMs share computational principles[69,70], but they are likely to implement these principles using radically different neural architectures[18,71]. Many state-of-the-art DLMs rely on transformers, a type of neural network architecture developed to solve the problem of sequence transduction. While current DLMs are an impressive engineering achievement, they are not biologically feasible, as 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 is using similar next-word prediction as an objective, it likely does so using a different implementation[69].

Classical versus deep language models

DLMs try to solve a fundamentally different problem than classical language models (CLMs). CLMs aim to uncover a set of generative (learned or innate) rules to be used in infinite, novel

situations, while considering extrapolation as the key principle needed for generating new sentences[72]. In contrast, DLMs aim to provide the appropriate linguistic output given the prior statistics of language use in similar contexts[17,18]. In other words, CLMs endeavor to describe observed language in terms of a succinct set of explanatory constructs while DLMs deemphasize interpretability and follow a performance-oriented predictive track, *learning* to produce passable linguistic outputs[73]. The internal contextual embedding space in DLMs can capture many aspects of the latent structure of human language, including the structure of syntactic trees, voice, co-references, and morphology, as well as long-range semantic and pragmatic dependencies[19,74,75]. "Direct fit"[18] to the data does not imply, however, that the network has learned to implement generative rules to be used in completely new contexts (out-of-distribution extrapolation). Rather, DLMs may simply rely on brute-force memorization and interpolation to learn how to generate the appropriate linguistic outputs in light of prior contexts[18].

Conclusion

Linguistics aims to expose the hidden underlying structure of language. The classical linguistics paradigm seeks a simple and interpretable set of linguistic rules, which can be used to explain the plethora of speech acts we produce in real-world contexts. DLMs, in contrast, provide an alternative self-supervised learning approach, which can derive context-specific representations of language from the surface statistics of the way people use words in real life. While classical language models may be more elegant and interpretable, the family of DLMs may better capture the biological process of language acquisition and language production. Furthermore, DLMs may be better positioned to deal with the numerous contextual exceptions and violations of any given linguistic rule.

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"[76]. 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 not sufficient 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[77]. 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, Liat Hasenfratz, Rita Goldstein, Sebastian Michelmann, Meir Meshulam, Manoj Kumar, Roi Reichart, Malcolm Slaney for technical and conceptual assistance that motivated and informed the writing of this manuscript. 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

Stimulus and transcription
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 in order to improve the accuracy of the alignment. The audio was downsampled to 11 kHz and the Penn Phonetics Lab Forced Aligner was used to automatically align the audio to the transcript[78]. 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 each segment (containing about 830 words) as the story unfolds. The first group was exposed to the first two words in the story, and were asked to predict the upcoming (i.e., third) word in the story. After entering their prediction, the actual next word in the story was revealed, and participants were asked again to predict the upcoming next (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 group provided a prediction for each word in the first segment of the story. 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 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, in accordance with 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. Thus, in the main paper we mainly focus on the left hemisphere, but for completion we also present maps for the right hemisphere in supplementary figures. 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 the duration of invasive monitoring 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 electrode relative to the cortical surface was determined from co-registered MRIs or CTs following the procedure described by Yang and colleagues[79]. 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[80], 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, then took the logarithm of each power time course estimate. These estimates were z-scored, and then averaged across these frequencies in order 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[80,81] or an ICA-based method[82] (based on the noise profile of the participant). High-frequency broadband (HFBB) power frequency provided evidence for high positive correlation between local neural firing rates and high gamma activity[83]. That is, the high gamma band fluctuation exhibited good estimations in the neural spiking population near each electrode[84]. 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[85].

Electrode-wise forward encoding model over time

The goal of this analysis was to use word embeddings to predict held-out neural data for individual electrodes (Fig. 2B). We used three sets of vectorial representation of words: 1) arbitrary embeddings sampled from uniform (-1,1) distribution; 2) GloVe embeddings; 3) GPT2 contextual embeddings. These embbedings were used to predict neural data in each electrode separately at varying time points relative to word onset. For each time point, we averaged across a 200 ms window. We assessed the performance of these models in predicting neural responses for held-out data using a 10-fold cross-validation procedure. The neural data (HFBB ECoG responses to each word) 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 and then applied those weights to predict the HFBB responses to the other 10% of the words (Fig. 2B). 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

In order to identify significant electrodes we used a randomization procedure. At each iteration, we randomized the phase of the signal of each of the electrodes, 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, for each electrode, the maximal value of the encoding model across all 161 lags was retained. 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) was computed as the percentile of the maximum value of the non-permuted encoding model 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 the multiple electrodes we used false-detection-rate (FDR[86]). In order to statistically assess the difference between the performance of two encoding models for the same electrode at specific lags (Fig. 4A), we subtracted the values of the two encoding models 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 resulting p-values to control the false discovery rate[FDR; 86].

To test the significance for each lag for the average encoding plots (Fig. 4B-D, S2 and S5) we used a bootstrap hypothesis test to compute a *p*-value for each lag[87]. 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[87]. To account for multiple tests across lags, we adjusted the resulting *p*-values to control the false discovery rate[FDR; 86]. A threshold was chosen to control the FDR at .01. We used a permutation test to assess significant differences in average encoding (Fig. 4B-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.

Decoding model over time

The goal of this analysis was to predict words from the neural signal. The input neural data was 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 the embedding of a word from the neural signal at a specific lag. The data was split into 5 stratified folds and used in a cross-validation procedure. The stratified folds preserve the percentage of instances of each class. Each fold consisted of a mean of 722.72 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.

In order 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 an ROC-AUC for the label. We plotted the weighted ROC-AUC according to the frequency of the word in the test set (which was equal to the frequency in the training set due to the stratified split). We chose words that had at least 5 repetitions in the training set (74% of the overall words in the narrative; see Appendix I for word list).

In order to improve the performance of the decoder, we implemented an ensemble of models. For each lag, we independently trained 10 decoders with randomized weight initializations and randomized the batch order fed into the neural net. 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

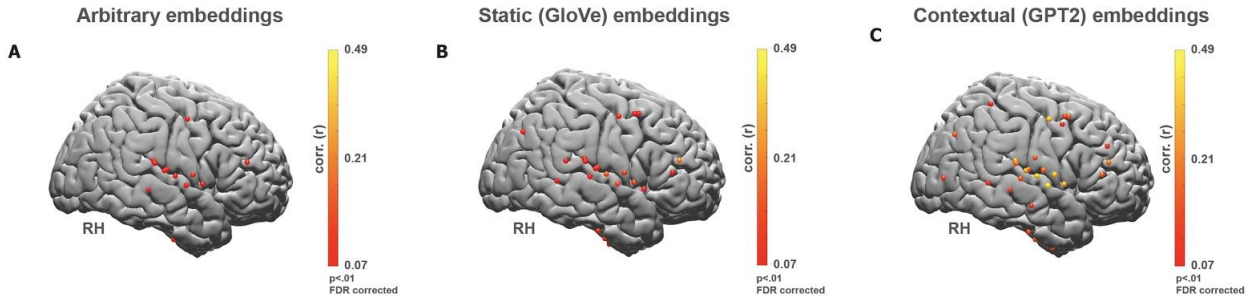


Figure S1. Right hemisphere encoding results also show advantage for contextual (GPT2) embeddings over static (GloVe) and arbitrary embeddings, Right Hemisphere maps for correlation between **A)** predicted and actual word responses for the arbitrary embeddings (nonparametric permutation test; $p < .01$,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

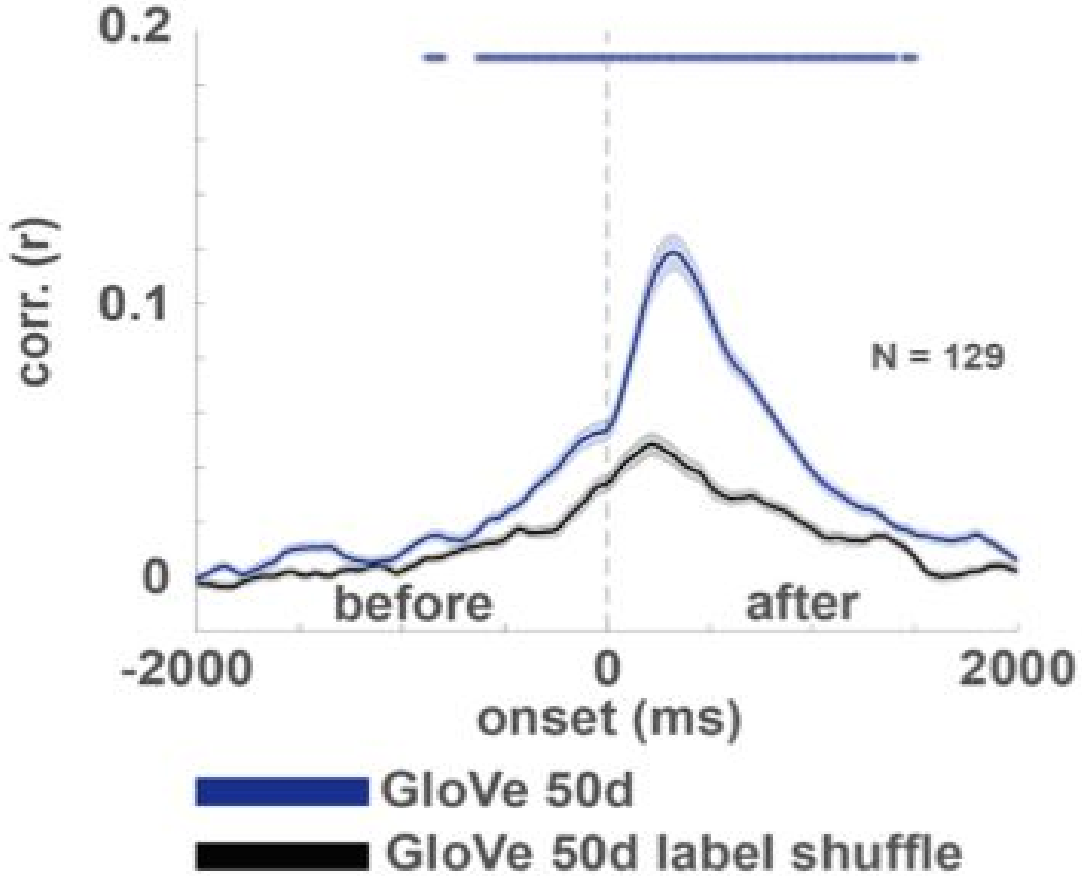
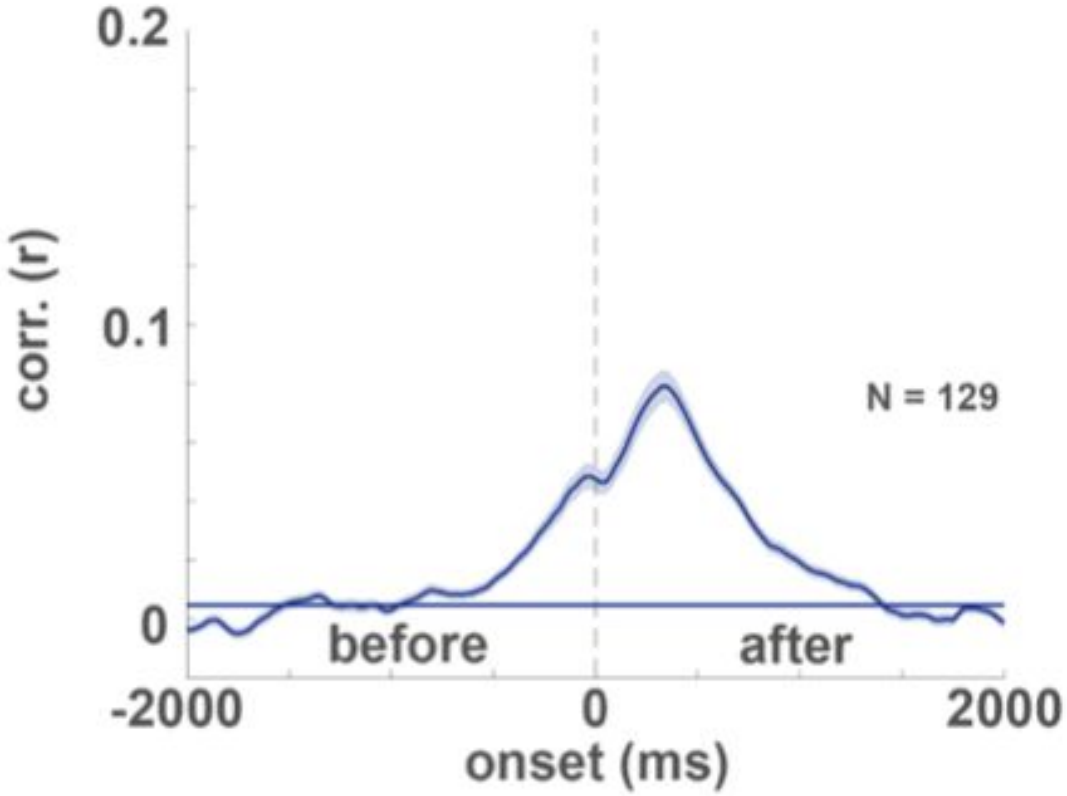


Figure S2. GloVe's space embedding attributes. It can be argued that GloVe based encoding outperforms arbitrary based encoding due to a general property of the space that GloVe embeddings induce (for example they are closer / further away from each other). In order to control for this possible confound 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 S3. 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. In order to make sure that the signal predicted before the onset is not a result of indirect correlation, we regressed out the embedding of the previous word from each word and re-run the encoding analysis. This also yields 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 co-occurrence of the words.

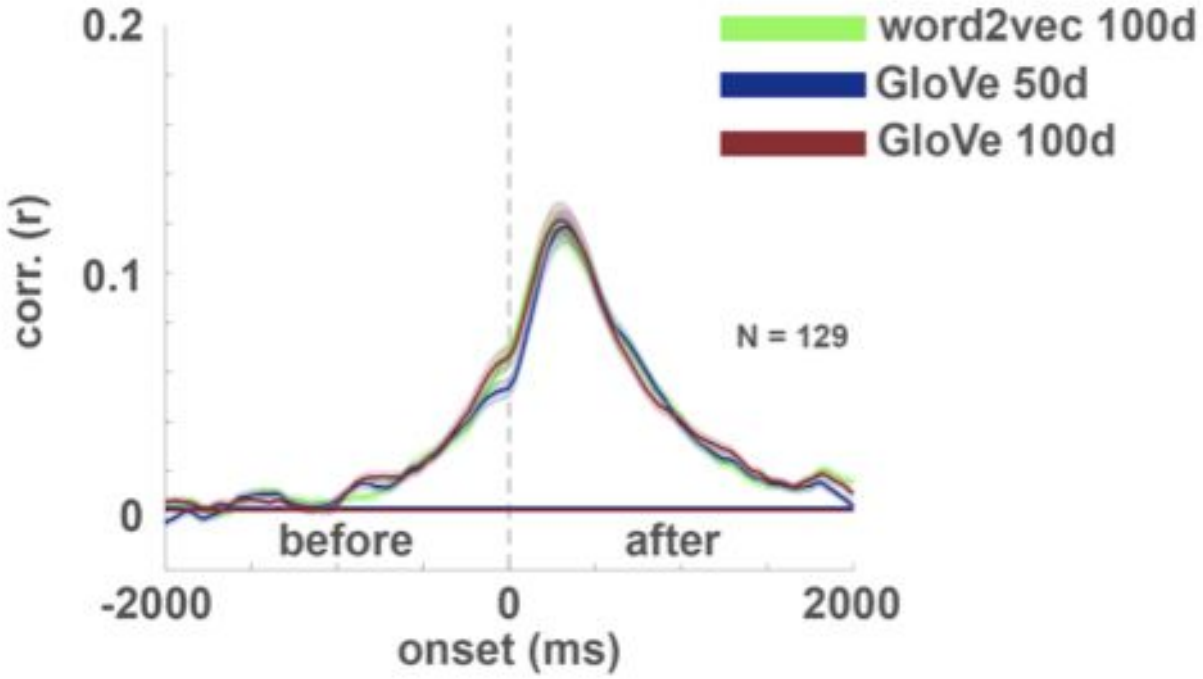


Figure S4. Comparison of GloVe- and word2vec-based static embeddings. The encoding procedure was repeated for two additional static embeddings using the electrodes that were found significant for GloVe-50 encoding on the left hemisphere (Fig. 3B). Each line indicates the encoding model performance averaged 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.

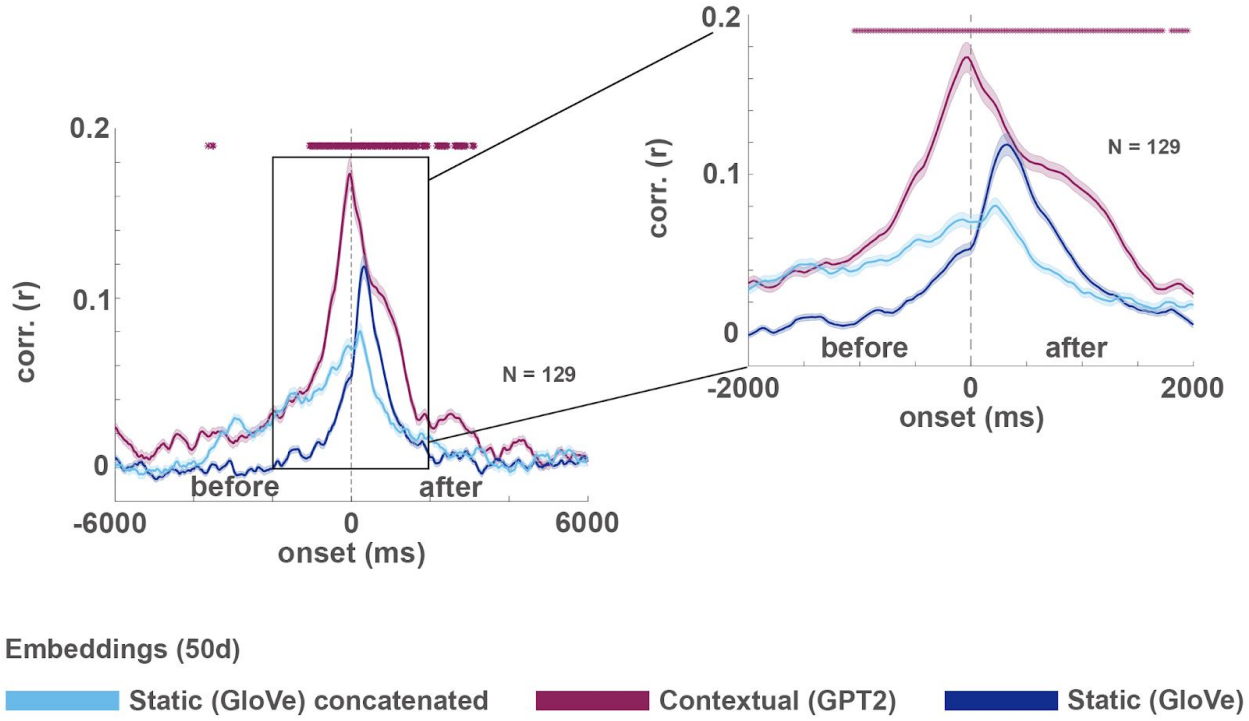


Figure S5. Comparison of GPT2 and concatenation of static embeddings. The increased performance of GPT2 based contextual embeddings encoding may be attributed to the fact that it consists of information about the identity of the previous words. In order to examine this possibility, we concatenated 5 GloVe words and reduced their dimensionality to 50 features. Still, GPT2 based encoding outperformed the mere concatenation prior to word onset, suggesting that GPT2's ability to compress the contextual information improves the ability to model the neural signals before word onset.

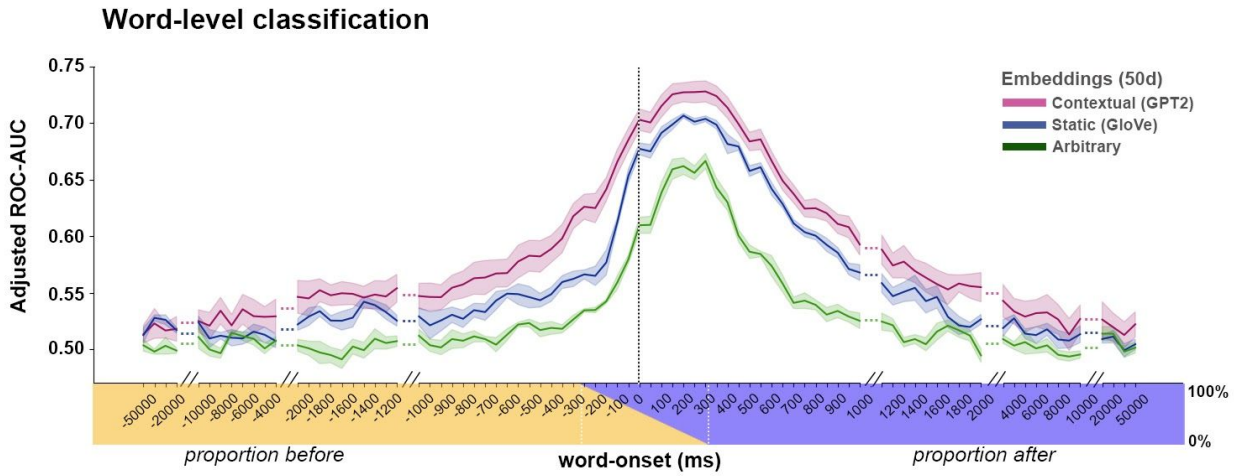


Figure S6. Decoding model adjusted for differences in baseline by subtracting the misaligned thresholds in Figure 7. *Word-level classification* for contextual embeddings (GPT2; purple), static embeddings (GloVe; blue), and arbitrary embeddings (green), after the subtraction of the corresponding misaligned baseline from each embedding. The units of Adjusted ROC-AUC are the same as ROC-AUC but the baseline (0.5) is adjusted according to the corresponding baseline of each set of embeddings. The *x*-axis labels indicate the center of each 625 ms window used for decoding at each lag (between -60 to 60 sec). The colored strip indicates the proportion of pre- (yellow) and post- (blue) word onset time points contained 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. Note that the adjustment for the uneven baselines did not change the results.

Appendix I - Word List

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

Appendix II - Model Details

Layer (type)	**Output Shape**	**Param #**
input_1 (InputLayer)	[(10, 107)]	0
conv1d (Conv1D)	(8, 160)	55680
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		**219,670**
Trainable parameters		**218,710**
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. [88]

1. Hacken, P. T. & Ten Hacken, P. Andrew Radford. Syntactic Theory and the Structure of English: A minimalist approach. Cambridge University Press, 1997. £18.95, ISBN 0-521-47707-7. Andrew Radford. Syntax: A minimalist introduction. Cambridge University Press, 1997. £14.95, ISBN 0-521-58914-2. *Natural Language Engineering* vol. 7 87–97 (2001).
2. Whaley, L. J. Mark C. Baker, The atoms of language: the mind's hidden rules of grammar. Oxford: Oxford University Press, 2001. Pp. xi 276. *Journal of Linguistics* vol. 39 699–700 (2003).
3. 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).
4. The 'Five Graces Group' *et al.* Language Is a Complex Adaptive System: Position Paper. *Language Learning* vol. 59 1–26 (2009).
5. Goldberg, A. E. *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. (Princeton University Press, 2019).
6. Lewis, M. *et al.* BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv [cs.CL]* (2019).
7. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv [cs.CL]* (2019).
8. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* (2018).
9. 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).

10. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003).
11. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. (2018).
12. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* **1**, (2019).
13. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv [cs.CL]* (2020).
14. Rosset, C. Turing-nlg: A 17-billion-parameter language model by microsoft. *Microsoft Blog* (2019).
15. Zaheer, M. *et al.* Big Bird: Transformers for Longer Sequences. *arXiv [cs.LG]* (2020).
16. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J. & Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. in *28th USENIX Security Symposium* 267–284 (2019).
17. Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L. & Lewis, M. Generalization through Memorization: Nearest Neighbor Language Models. *arXiv [cs.CL]* (2019).
18. 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).
19. Linzen, T. & Baroni, M. Syntactic Structure from Deep Learning. *Annu. Rev. Linguist.* (2021) doi:10.1146/annurev-linguistics-032020-051035.
20. Chomsky, N. *Knowledge of Language: Its Nature, Origin, and Use*. (Greenwood Publishing Group, 1986).

21. Pinker, S. & Ullman, M. T. The past and future of the past tense. *Trends Cogn. Sci.* **6**, 456–463 (2002).

22. Pinker, S. Rule of language. *Science* **253**, 530–534 (1997).

23. 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.

24. Pereira, F. *et al.* Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).

25. 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).

26. 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).

27. 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.

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

29. Donhauser, P. W. & Baillet, S. Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron* **105**, 385–393.e9 (2020).

30. 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.

31. Jain, S. & Huth, A. G. Incorporating Context into Language Encoding Models for fMRI. doi:10.1101/327601.

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

33. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8619–8624 (2014).

34. Kriegeskorte, N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu Rev Vis Sci* **1**, 417–446 (2015).

35. Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J. & Kurth-Nelson, Z. Deep Reinforcement Learning and Its Neuroscientific Implications. *Neuron* **107**, 603–616 (2020).

36. Clark, K., Khandelwal, U., Omer, L. & D, M. C. Inducing brain-relevant bias in natural language processing models. *arXiv* (2019).

37. Taylor, W. L. 'Cloze Procedure': A New Tool for Measuring Readability. *Journal. Q.* **30**, 415–433 (1953).

38. 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).

39. 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).

40. 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).

41. 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).
42. 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).
43. Beim Graben, P., Gerth, S. & Vasishth, S. Towards dynamical system models of language-related brain potentials. *Cogn. Neurodyn.* **2**, 229–255 (2008).
44. 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).
45. Kutas, M. & Hillyard, S. A. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* **207**, 203–205 (1980).
46. Kutas, M. & Hillyard, S. A. Brain potentials during reading reflect word expectancy and semantic association. *Nature* **307**, 161–163 (1984).
47. 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).
48. Chivvis & Dana. "So a Monkey and a Horse Walk Into a Bar". (2017).
49. 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).

50. 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).
51. 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).
52. Athiwaratkun, B., Wilson, A. & Anandkumar, A. Probabilistic FastText for Multi-Sense Word Embeddings. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018) doi:10.18653/v1/p18-1001.
53. Lan, Z. *et al.* ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv [cs.CL]* (2019).
54. Clark, K., Luong, M.-T., Le, Q. V. & Manning, C. D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv [cs.CL]* (2020).
55. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *Neuroimage* **56**, 400–410 (2011).
56. Mandrekar, J. N. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology* vol. 5 1315–1316 (2010).
57. 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.
58. Kuperberg, G. R. & Jaeger, T. F. What do we mean by prediction in language comprehension? *Lang Cogn Neurosci* **31**, 32–59 (2016).
59. Hart, B. & Risley, T. R. Meaningful differences in the everyday experience of young

American children. **268**, (1995).

60. Weisleder, A. & Fernald, A. Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychol. Sci.* **24**, 2143–2152 (2013).
61. 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).
62. Tan, H. & Bansal, M. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. *arXiv* (2020).
63. 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).
64. Yeshurun, Y. *et al.* Same Story, Different Story: The Neural Representation of Interpretive Frameworks. *Psychol. Sci.* **28**, 307–319 (2017).
65. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948).
66. Shannon, C. E. Prediction and Entropy of Printed English. *Bell System Technical Journal* vol. 30 50–64 (1951).
67. 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).
68. 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.
69. Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).

70. 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.

71. Heeger, D. J. Theory of cortical function. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1773–1782 (2017).

72. Chomsky, N. ASPECTS OF THE THEORY OF SYNTAX. (1964) doi:10.21236/ad0616323.

73. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16**, 199–231 (2001).

74. 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.

75. 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.

76. Chomsky, N. Syntactic Structures. (1957) doi:10.1515/9783112316009.

77. Marcus, G. F. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. (MIT Press, 2019).

78. Yuan, J. & Liberman, M. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* **123**, 3878 (2008).

79. Yang, A. I. *et al.* Localization of dense intracranial electrode arrays using magnetic resonance imaging. *NeuroImage* vol. 63 157–165 (2012).

80. 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).

81. Lachaux, J. P., Rudrauf, D. & Kahane, P. Intracranial EEG and human brain mapping. *J. Physiol. Paris* **97**, 613–628 (2003).

82. Michelmann, S. *et al.* Data-driven re-referencing of intracranial EEG based on independent component analysis (ICA). *J. Neurosci. Methods* **307**, 125–137 (2018).

83. Jia, X., Tanabe, S. & Kohn, A. Gamma and the Coordination of Spiking Activity in Early Visual Cortex. *Neuron* **77**, 762–774 (2013).

84. 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).

85. Honey, C. J. *et al.* Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423–434 (2012).

86. 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).

87. Hall, P. & Wilson, S. R. Two Guidelines for Bootstrap Hypothesis Testing. *Biometrics* vol. 47 757 (1991).

88. 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.