Exploring Temporal Sensitivity in the Brain Using Multi-timescale Language Models: An EEG Decoding Study

Sijie Ling
University of Alberta
Department of Psychology
Alberta Machine Intelligence Institute
sling1@ualberta.ca

Alex Murphy
University of Alberta
Department of Computing Science
Alberta Machine Intelligence Institute
amurphy3@ualberta.ca

Alona Fyshe
University of Alberta
Department of Computing Science
Department of Psychology
Alberta Machine Intelligence Institute
alona@ualberta.ca

The brain's ability to perform complex computations at varying timescales is crucial, ranging from understanding single words to grasping the overarching narrative of a story. Recently, multi-timescale long short-term memory (MT-LSTM) models (Mahto et al. 2020; Jain et al. 2020) have been introduced, which use temporally tuned parameters to induce sensitivity to different timescales of language processing (i.e., related to near/distant words). However, there has not been an exploration of the relationship between such temporally tuned information processing in MT-LSTMs and the brain's processing of language using high temporal resolution recording modalities, such as electroencephalography (EEG).

To bridge this gap, we used an EEG dataset recorded while participants listened to Chapter 1 of "Alice in Wonderland" and trained ridge regression models to predict the temporally tuned MT-LSTM embeddings from EEG responses. Our analysis reveals that EEG signals can be used to predict MT-LSTM embeddings across various timescales. For longer timescales, our models produced accurate predictions within an extended time window of ±2 s around word onset,

Action Editors: Marianna Apidianaki, Abdellah Fourtassi, and Sebastian Padó. Submission received: 17 December 2023; revised version received: 30 April 2024; accepted for publication: 19 June 2024.

https://doi.org/10.1162/coli_a_00533

while for shorter timescales, significant predictions are confined to a narrower window ranging from -180 ms to 790 ms. Intriguingly, we observed that short timescale information is not only processed in the vicinity of word onset but also at more distant time points.

These observations underscore the parallels and discrepancies between computational models and the neural mechanisms of the brain. As word embeddings are used more as in silico models of semantic representation in the brain, a more explicit consideration of timescale-dependent processing enables more targeted explorations of language processing in humans and machines.

1. Introduction

Language, at its core, is a hierarchical system in which smaller units combine into larger units. In spoken language processing, phones combine into phonemes, which combine into meaningful morphemes/words, which further combine into phrases, sentences, paragraphs, and passages (and similarly for written language). Psycholinguists have studied the human mind to define the mechanistic processes that underlie linguistic processing at multiple timescales: words, sentences, and longer-range contexts (Traxler 2011). Brain imaging has shown that the brain tracks language processing via numerous cognitive mechanisms (Ding et al. 2016; Lerner et al. 2011), which relate hierarchical language processing to different *timescales*.

Here, we also study the timescales of natural language and the corresponding processes in the human brain. However, unlike previous work, we use high temporal resolution electroencephalography (EEG) recordings alongside a data-driven method for defining timescales. When studying language in the brain, timescales have been operationalized in different ways. For example, in Jain and Huth (2018) and Chen et al. (2024), timescales are defined as the sensitivity of language models to words in specific ranges of prior context (i.e., 32–64, 64–128 words). Contextual representations have been used in prior work to study correspondences between linguistic timescales and functional magnetic resonance imaging (fMRI) responses (Mahto et al. 2020; Jain et al. 2020; Vo et al. 2023). In our analysis, we derive language representations from a model that preserves contextual information at different rates, and then relate those representations to human brain activity. We use a multi-timescale long short-term memory (MT-LSTM) model (Mahto et al. 2020; Jain et al. 2020), which uses a data-driven definition of timescales (see Section 2.3). The MT-LSTM has continuously parameterized timescales that change smoothly in order to increase sensitivity to information from the word level to the paragraph level and beyond. We partition these MT-LSTM units based on their sensitivity to (i) long, (ii) medium, and (iii) short timescales in a data-driven way. Our results add to previous fMRI findings by combining MT-LSTM representations with higher temporal resolution EEG recordings, which allow for a more fine-grained analysis of the timescales of language processing.

Our analysis aims to identify and characterize brain networks with temporal sensitivity corresponding to the timescales of language represented by the MT-LSTM. We use 12-minute EEG recordings (Bhattasali et al. 2020) collected while participants passively listened to the first chapter of *Alice in Wonderland*. We used a decoding framework, training L_2-regularized ridge regression models (Hoerl and Kennard 1970) to robustly predict MT-LSTM representations from EEG data. We used the temporal generalization method (TGM) (Dehaene and King 2016; Fyshe 2020) to track the temporal stability of the brain's timescale representations during word processing. We then performed a spatial analysis using sensor subsets (Rafidi 2018) to find the brain areas responsible for successful decoding of MT-LSTM representations. Our analysis reveals several insights into the brain's temporal sensitivity to the varying timescales of language. Specifically, we show that:

1. Information pertaining to different language timescales can be decoded from small windows (100 ms) of EEG signals.
2. Brain responses predict MT-LSTM embeddings both before and after word onset.
3. As the timescale increases, MT-LSTM embeddings can be decoded from EEG for a longer time window around word onset.
4. As shown in previous work (Huth et al. 2016; Lerner et al. 2011; Jain and Huth 2018), word representations can be localized throughout the brain, but at time points further from word onset, sensors over the temporal lobe and prefrontal cortex give best decoding performance.

To the best of our knowledge, ours is the first study to examine the connection between brain responses and timescale-tuned word embeddings using temporally rich neural recordings.

2. Background

2.1 Timescales in Human Language Comprehension

Temporal Receptive Windows (TRWs) describe the sensitivity of a neuron or a cortical microcircuit to a specific temporal window (Hasson et al. 2008), a concept that has been extended to study language in the brain. In studies of the brain's sensitivity to language timescales, stimuli are selected such that they exhibit a range of linguistic information requiring processing at multiple temporal levels (Xu et al. 2005; Lerner et al. 2011; Brennan and Pylkkänen 2012). Brain recordings related to specific timescales are then extracted and analyzed. Previous paradigms assume that text scrambled at one timescale will disrupt brain processes sensitive to the longer timescales, while leaving processes relating to the shorter timescales intact. For instance, random word lists activate sub-networks that process sounds and words, but not phrases or sentences (ten Oever et al. 2022). Based on this idea, Xu et al. (2005), Lerner et al. (2011), and Farbood et al. (2015) used fMRI to find that a hierarchy of brain areas corresponds to increasing TRWs, extending from early auditory cortices along the superior temporal gyrus up to the intraparietal sulcus, as well as highlighting the role of the frontal cortex in sentence and paragraph level comprehension. Blank and Fedorenko (2020) further argued that left inferior frontal and temporal language regions cannot distinguish timescales longer than sentences and do not have distinct stages for this hierarchy. Using a more temporally precise method, Brennan and Pylkkänen (2012) discovered that the difference between sentence and scrambled word lists in magnetoencephalography (MEG) for language-related brain areas are concentrated at 250–300 ms after word onset.

Another approach is to record neural oscillations in the brain and test for their correspondence to the timescales of the language processing hierarchy. Ding et al. (2016, 2017) devised specific linguistic stimuli to test this: every word is its own unit; every 2 words form a phrase; every 2 phrases form a sentence. When they fixed the word presentation rate to 4 Hz, they found 3 entrained frequencies from EEG and MEG that correspond to the word (1 Hz), phrase (2 Hz), and sentence (4 Hz) units in their linguistic stimuli. Keitel, Gross, and Kayser (2018), Kaufeld et al. (2020), and Meyer (2018) introduced the framework of oscillations for chunking (i.e., oscillations mark the edge of timescales), and Kazanina and Tavano (2023) discussed some potentially theoretical challengers (e.g., various wavelengths for the same linguistic unit).

Multi-timescale processing has also been proposed for other non-language brain functions, such as vision (Hasson et al. 2008; Honey et al. 2012; Murray et al. 2014; Zeraati et al. 2023), music perception (Farbood et al. 2015), memory (Gao et al. 2020), and decision-making processes (Spitmaan et al. 2020). Taken together, multi-timescale processing can be postulated as one of the brain's global organizing principles (Raut, Snyder, and Raichle 2020). Given that evidence shows linguistic representations can be found in many parts of the brain (Huth et al. 2016), the timescale structure of other cognitive functions may also be involved in language comprehension.

2.2 Encoding and Decoding Methods

Encoding and decoding methods (Mitchell et al. 2008; Huth et al. 2016) are one way to study the flow of information during cognitive processing. Such methods can help to resolve questions of linguistic representation in the brain by learning a mapping between an abstract language property and brain responses to stimuli with that property. A typical encoding / decoding framework consists of 4 steps:

1. High-dimensional vectors are derived to represent properties in the stimuli.
2. These vectors are paired to the brain responses recorded when a person processes the corresponding stimuli.
3. A model is trained to learn a mapping between a subset of pairs.
4. A metric (i.e., accuracy) is computed on held-out pairs in a test set.

In encoding, step 3 requires that the stimulus property vector is used to predict brain responses, while decoding involves the opposite direction, predicting the stimulus property vector from brain responses. We assume the common principle that successful decodability of a stimulus from neural data implies the presence of stimulus-relevant information in that signal (Reddy and Wehbe 2021; Kriegeskorte and Kievit 2013; la Tour et al. 2022). If the activity in some brain regions can predict stimulus properties well during a period of time, it indicates that the property (or a correlate of it) is processed in those brain regions during that time period.

Decoding models have also been used to explore language processing timescales. By mapping brain responses to linguistic representations, researchers have identified the contributions of various brain areas to corresponding timescales during comprehension of words (Mitchell et al. 2008; Huth et al. 2016), phrases (Fyshe et al. 2019), syntax and parts-of-speech (Hale et al. 2018; Murphy et al. 2022), sentence segments (Wehbe et al. 2014a), and complete sentences (Pereira et al. 2018). Although short timescales can be modeled by averaging or concatenating non-contextual word vectors for each of the constituent words, longer timescales require a more sophisticated approach.

In our work, we represent longer timescales using the MT-LSTM, a variant of the LSTM model. LSTMs are trained to simply predict the next word in a sequence, yet the models encode numerous additional linguistic features. Visualization and ablation studies have found that some neurons correspond to multi-level linguistic features, such as suffix starts, word endings, and long-range subject-verb dependency (Linzen, Dupoux, and Goldberg 2016; Gulordava et al. 2018; Kementchedjhieva and Lopez 2018; Lakretz et al. 2019; Chien and Honey 2020). Modifying the structure or parameters of the neurons to induce sensitivity to long timescale information can improve model performance (Lin et al. 2015; Hwang and Sung 2017; Singh and Lee 2017; Shen et al. 2018; Mahto et al. 2020). For larger models like GPT (Radford et al. 2018) and BERT (Devlin et al. 2019), their layered structure shows a hierarchy of timescales (Goldstein et al. 2022a), and linear filters can be applied on the representation to separate timescale information (Chen et al. 2024). Timescales have been extracted from RNNs (Chien and Honey 2020), LSTMs (Jain and Huth 2018; Jain et al. 2020), GPT (Caucheteux, Gramfort, and King 2021; Goldstein et al. 2022b; Heilbron et al. 2022), and BERT (Chen et al. 2024) and compared with brain responses to language. Brain areas found to be sensitive to different timescales are consistent with neurolinguistic analyses of scrambled text processing.

One limitation of the aforementioned studies is that most use fMRI recordings to analyze timescale representations. Because normal speech is faster than the time required to acquire a full fMRI volume ($\sim$ 2 seconds), word representations need to be downsampled to correspond to the sampling rate of fMRI. This results in a loss of information and a temporally imprecise picture of brain processing. This hinders our ability to address certain linguistic questions with temporal precision—for example, determining the function of neural oscillations in processing linguistic elements of different timescales (Kazanina and Tavano 2023).

2.3 MT-LSTM Model

To derive timescale-sensitive linguistic representations, we used the MT-LSTM model proposed in Mahto et al. (2020). This model is smaller in scale than the more recent Transformer models, which makes it easier to train, especially in cases where limited linguistic materials exist to train and validate the model. The MT-LSTM model has a distribution of timescales allowing the model to take into account the time-varying characteristics of natural language. The timescales in the model are defined parametrically, and the model is induced to capture information carried over by these timescales for prediction. The embeddings from various timescales can be extracted directly from the model in order to derive a mapping from brain responses to these embeddings.

Tallec and Ollivier (2018) discovered that the persistence of information in the LSTM contextual representations follows an exponential decay curve, and Mahto et al. (2020) further solidified this multi-timescale property by modifying the forget gate bias b_f of individual LSTM neurons. A neuron with a more negative b_f tends to up-weight new inputs to the LSTM cell state and thereby model shorter timescales. In contrast, a more positive b_f results in the neuron preserving cell state from prior tokens over newly processed inputs, thereby resulting in sensitivity to longer timescales. Based on the decay rate of information, the timescale T, or **forgetting time** of an LSTM neuron can be computed as:

$$T = \frac{1}{\log(1 + e^{-b_f})} \tag{1}$$

This formula means that we can induce sensitivity to various timescales by modulating the forget gate biases with timescale-specific values of T.

Besides the explicit definition of timescale, the MT-LSTM model has 2 other advantages. First, the MT-LSTM is inspired by a property of natural language: The mutual information between tokens follows a power law decay as distance increases (Lin and Tegmark 2016). However, the decay rate in LSTM units is exponential. To make exponential LSTM units fit the observed power law, the MT-LSTM timescales are altered to follow an inverse gamma distribution. Mahto et al. (2020) tuned the decay parameter to optimize the accuracy of next token prediction. The best parameter setting results in a range of timescales from 0.17 words to 360k words.

Second, MT-LSTM units have been shown to correspond with linguistic features. Vo et al. (2023) probed the MT-LSTM model to detect the extent to which timescale-related information in the model was related to various linguistic features. For example, part-of-speech is found to be best represented at word and sentence-level timescales, while document-level information (e.g., topics) corresponds best to paragraph and multi-paragraph timescales.

In summary, the embeddings of different timescales in the MT-LSTM model can be used as a proxy for language representations at various timescales. These timescales are explicitly defined and past work has shown that different types of linguistic information are represented at each timescale.

3. Methods

3.1 EEG Data

We used the *Alice* dataset (Bhattasali et al. 2020), a public dataset containing EEG signals collected while participants ($n = 49$) passively listened to *Alice's Adventure in Wonderland* Chapter 1. The EEG data were recorded at 500 Hz with 59 channels (reference channels excluded) using the Easy-M10 montage. Each recording session lasted for 12.4 minutes and contains 2,129 labeled word tokens. We applied further preprocessing (see Appendix A) to the data to remove artifacts and chose 19 participants (3 male, age range 18–25 years) for the final analysis. Participant exclusion criteria are given in Appendix A. A temporal window of $[-2, 4]$ seconds around the onset of each word of interest was extracted and linear detrending was applied on the epoched EEG data.

3.2 MT-LSTM Embeddings

Model Structure. Our MT-LSTM model follows Jain et al. (2020), who use a stateful LSTM with three layers. This model takes in a 400-dimension word embedding and outputs a vector of the same dimension. The first two layers each have 1,150 neurons, and the third has 400 neurons. Before the LSTM layers, the model has an embedding layer, which encodes each word token in the vocabulary from an index value into a 400-dimension word embedding. The same mapping is used to convert the 400-dimension output of the LSTM into a probability distribution over possible next tokens in the vocabulary, from which prediction loss and perplexity can be calculated.

Model Parameters. In Layer 1 of the LSTM, immediately after the input embedding layer, half of the 1,150 neurons are assigned a timescale of $T = 3$ and half with $T = 4$. These timescales ensure Layer 1 only processes short timescale information. In Layer 2, the timescales of the 1,150 neurons follow an inverse gamma distribution with $\alpha = 0.56$

and $\beta = 1$. In both Layer 1 and Layer 2, the forget gate bias b_f of each neuron is set to its assigned timescale according to the following equation

$$b_f = -\log(e^{\frac{1}{T}} - 1) \tag{2}$$

which is derived from Equation (1). The input gate biases b_i are set to be the negated value of the forget gate biases (i.e., $b_i = -b_f$) so that timescale sensitivity is controlled solely by the settings of the forget gate parameters. All other parameters, including those in Layer 3, are randomly initialized from a uniform distribution between [−0.1, 0.1] and optimized during training.

Training. We trained the MT-LSTM with Stochastic Gradient Descent (Amari 1993). In the pretraining stage, we used the Wikitext-2 dataset (Merity et al. 2017) (following Mahto et al. 2020) to train the MT-LSTM model for 1,600 epochs.

Fine-tuning. For fine-tuning, we constructed a Lewis-Carroll Set (LC set) with 3 books, *Alice's Adventures in Wonderland* (*"Alice in Wonderland"*), *Through the Looking-Glass*, and *What Alice Found There*, and additionally *Sylvie and Bruno* from Project Gutenberg. The data were parsed with the spaCy toolbox (Honnibal and Montani 2017) and all punctuation tokens were removed. We selected Chapter 1 of *Alice in Wonderland* (stimuli of *Alice* dataset) to be the test set (2.1K tokens) because EEG recordings were collected for this portion of the dataset. Chapters 2–3 were selected as the validation set (3.9K tokens) because they are likely the most similar to Chapter 1, and thus most representative for the chosen test set. All other text was selected to be training data (115K tokens) to build the token vocabulary. The vocabulary contained 7,006 unique tokens and covered 2,052 of the 2,129 tokens (867 of the 914 lexical tokens) in the test set.

To train using the LC dataset, we modified the model vocabulary to match this dataset and fine-tuned the model for another 20,000 epochs. The training results for multiple models were stable with average perplexity 83.57± 0.82 for the test set. We chose the model with smallest perplexity on the validation data to produce timescale-tuned embeddings.

3.3 Token Selection

In order to produce multi-timescale contextual word embeddings, we used the MT-LSTM model embeddings to represent w_t. After inputting token w_{t-1}, we recorded v_t, the hidden states of layer 2 as shown in Figure 1. Each v_t has a dimensionality of 1,150. Because we are interested in how semantic representations of context in the model predict the next token, we omit function words (e.g., *and*, *to*, *it*) from the decoding analysis, though they were still used during the training of the MT-LSTM model. We chose 800 from the 914 lexical words in the test set that were:

- Not at the beginning of the story or the end of the story, in order to decrease the EEG artifacts caused by inadaptation to the experiment at the beginning, and fatigue at the end.
- Not proper nouns (e.g., names), as such tokens have limited coverage in small datasets and are rarely updated from random initialization, resulting in a poorer semantic representation.

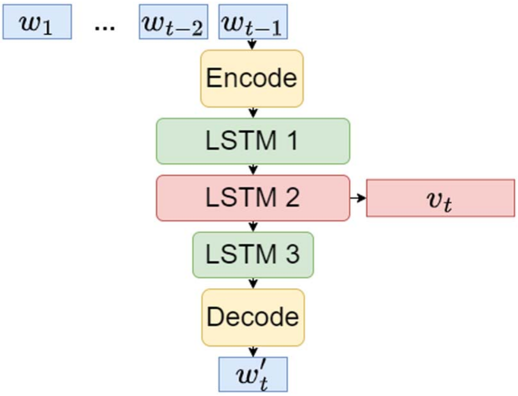


Figure 1
The diagram for the MT-LSTM model. The model has 3 LSTM layers. The encoder and the decoder share weights. To generate the representation for word w_t, we sequentially supply word w_1 through w_{t-1} to the model, from which the model produces a prediction for w'_t. For our analyses, the representation for word w_t is the output of the second LSTM layer (v_t), which is extracted after the model processes w_{t-1}.

- Not following the "unknown" token (<UNK>)[1] because the <UNK> token amalgamates semantic information over all unknown words, which will negatively affect the representation for the word of interest.

3.4 Decoding Method

Because we want to analyze the EEG's correspondence with each of the 1,150 MT-LSTM dimensions, we trained a decoding model to predict word embeddings from the EEG data. We then divided the embedding dimensions into groups based on their behavior during text processing (see Section 4.2) and measured the average decodability of each group. Averaging decodability improves signal-to-noise ratio (SNR), giving a better picture of the relation between MT-LSTM timescales and the brain's processing of language.

EEG data have high temporal resolution, allowing us to train multiple decoding models using different time windows of EEG data to produce a timeline of decoding performance. We selected 100 ms sliding windows (50 time steps) with a shift of 10 ms (5 time steps) and trained a separate decoding model with each window of data. High decoding performance in a specific temporal window implies that the brain responses at that time are sensitive to the corresponding timescale-tuned word embeddings.

3.5 Ridge Regression Model

Our training dataset consists of EEG windows (input) and word embeddings (target) pairs, $D = \{X_i, Y_{i,j}\}$, where $X_i \in \mathbb{R}^{2950}$ represents one flattened EEG window (50 time steps × 59 electrodes) and $Y_{i,j} \in \mathbb{R}$ is a scalar that represents the jth dimension of the ith word embedding. We followed common convention and used a L_2-regularized (Ridge)

[1] If a word is presented to the model that was not in the training vocabulary, it is assigned the label of <UNK> (unknown).

regression model to learn a mapping between EEG responses and word embeddings. We learn a linear mapping $W_j \in \mathbb{R}^{2950\times 1}$ using Equation (3), independently for each word embedding dimension ($N = 1{,}150$). Collecting the models for each dimension of the word embeddings into a matrix, we can specify the calculation as the following, where λ_j represents the ridge penalty for dimension j.

$$W_j = \underset{W_j}{\arg\min}\{\left\|Y_{i,j} - X_i W_j\right\|^2 + \lambda_j \left\|W_j\right\|^2\} \tag{3}$$

A separate model W_j was independently trained and evaluated for each dimension j of the word embeddings. A depiction of this training procedure is given in Figure 2.

3.6 Training and Evaluation Paradigm

We evaluated the model using 10-fold *nested* cross-validation. The cross-validation was repeated 10 times with distinct random partitions of data. Using the predictions from each of the 10 × 10 folds we calculated the Pearson correlation between the true and predicted MT-LSTM embedding and reported the average correlation over all folds.

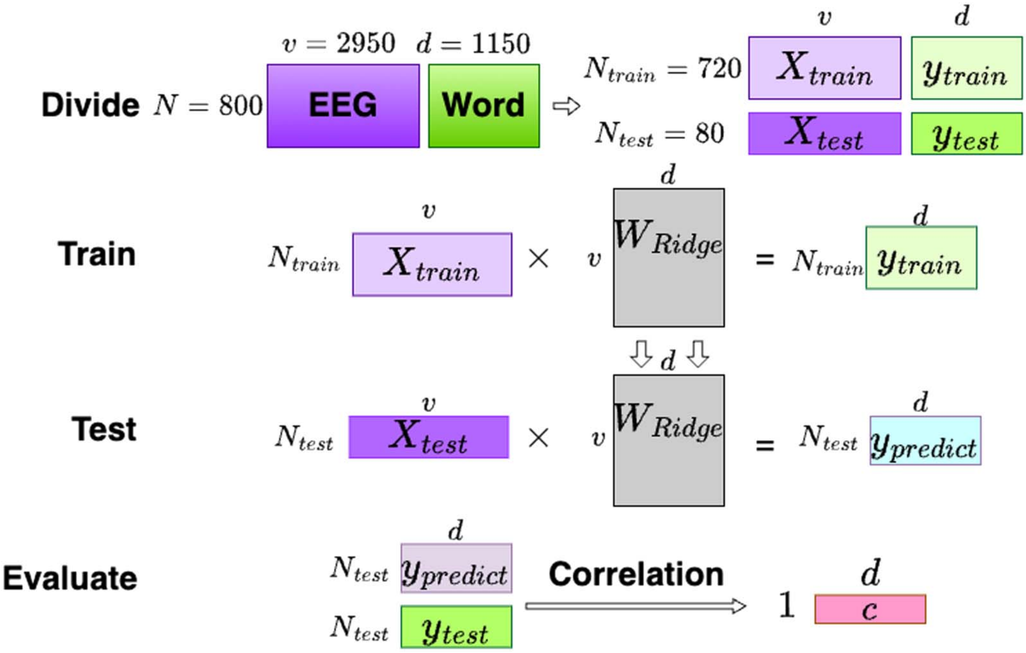


Figure 2
An example of one training fold, evaluating and testing a decoding model on one time window. The numbers in the figure show the dimension of each matrix. Divide: Data are split into train and test sets of size 720 and 80, respectively. EEG data are shown in purple, word embeddings are green. Train: We train a ridge regression model (W_{Ridge}, gray) to predict word embeddings from EEG. For brevity, we omit the step of independent regularization for each dimension. Test: W_{Ridge} is applied to test (hold-out) EEG data to predict the corresponding MT-LSTM embeddings. Evaluate: For each dimension, a correlation value (c, pink) is calculated between predicted embeddings ($y_{predict}$, blue) and true embeddings (y_{test}, green).

Figure 2 gives an overview of the procedure for one cross-validation fold. Data were divided into a training and test set; EEG data were normalized to have mean zero and standard deviation of one. We used Leave-One-Out Cross-Validation to select the best ridge regression hyperparameter (λ) during training. We computed decoding performance for every λ, and used the best performing λ to train one model on all the samples in the *training* set. The model was evaluated on the test set defined during the 10-fold cross validation process and we computed the Pearson correlation between true and predicted embeddings.

3.7 Temporal Generalization Method (TGM)

In the above analyses, for each time window of EEG signals, we trained a separate ridge regression model to predict individual dimensions of the MT-LSTM embedding. However, that performance curve does not reflect whether the underlying regression models at different time points are leveraging similar patterns in the EEG data. We used TGM (Dehaene and King 2016; Fyshe 2020) to address this question.

Instead of using training and test samples from the same time window, the TGM evaluates model predictions by testing samples from all time windows with respect to a fixed training window. With N time windows, the TGM produces a $N \times N$ dimension matrix that illustrates how stable different temporal windows are with regard to the learned model. If models trained on one time window can be generalized to another time window, the brain's language representations can be inferred to be similar. Conversely, if models trained on different windows have good individual performance but cannot generalize to other windows, it suggests a change in the brain's language representations across time.

3.8 Sensor-Subset Analysis

This analysis was used to determine which brain areas contribute most to the performance of the decoding model. The training process was the same as in Section 3.6, but we used a subset of EEG sensors to train the model. For this analysis, a sensor subset consists of a central sensor and all adjacent sensors according to the Easycap-M10 montage. The sensor subsets provide more stable results for comparing the contribution of different brain areas by diminishing the influence of noise in a single channel. For each specified timescale of interest, we plotted the test correlation for each sensor on the scalp topography with the MNE toolbox (Gramfort et al. 2013).

3.9 Significance Testing

To assess if decoding results are significantly above chance, we performed a permutation test analysis. Specifically, for 100 iterations, we shuffled the rows of the MT-LSTM embedding matrix such that all connection between word identity and associated embedding is removed. For each of our proposed analyses, we ran permutation tests with different random seeds and followed the same analysis procedure as for the non-permuted data. We used kernel density estimation with a Gaussian kernel (Chen 2017) to generate a null hypothesis distribution from 100 permutation test results and used the null distribution to calculate a p-value for the non-permuted assignment. We used the Benjamini-Hochberg-Yekutieli False Discovery Rate (FDR) correction (Benjamini and Hochberg 1995) with a family-wise error rate of $\alpha = .05$ to correct for multiple comparisons.

4. Results

4.1 Analysis 1: Decoding MT-LSTM Embeddings

We used the experimental procedure in Section 3.6 to explore whether EEG can predict the contextual MT-LSTM embeddings, considering all timescales at once. Recall that, for each word, the MT-LSTM hidden state v_t is the intermediate output of predicting the next word w_t (see Figure 1). Therefore v_t contains some semantic information about word w_t. We expected that the MT-LSTM embeddings would be decoded from EEG after word onset. Additionally, the MT-LSTM model has processed words w_1 through w_{t-1}, so it has captured some essence of the word's context, reflected in v_t. This contextual information is expected to be processed by the brain in the vicinity of w_t, so we also expect to observe accurate decoding performance before word onset due to this overlap in information.

The result for decoding MT-LSTM embeddings is shown in Figure 3. As expected, the peak decoding performance appears around word onset. The range is from 170 ms before the onset to 700 ms after onset. This range is wider than the [0, 400] ms period reported in Wehbe et al. (2014b), who used word embedding vectors to predict MEG signals (i.e., encoding) during a controlled reading paradigm. This indicates that the EEG recordings have captured contextual representations that are predictive of the MT-LSTM embeddings. We also found above chance decoding performance at time points distant from word onsets [−2, −1, 1.2, 2.1, 4] s. This suggests that the more abundant contextual information in the MT-LSTM embeddings correlates with EEG signals far from the onset of the word of interest, during continuous speech processing.

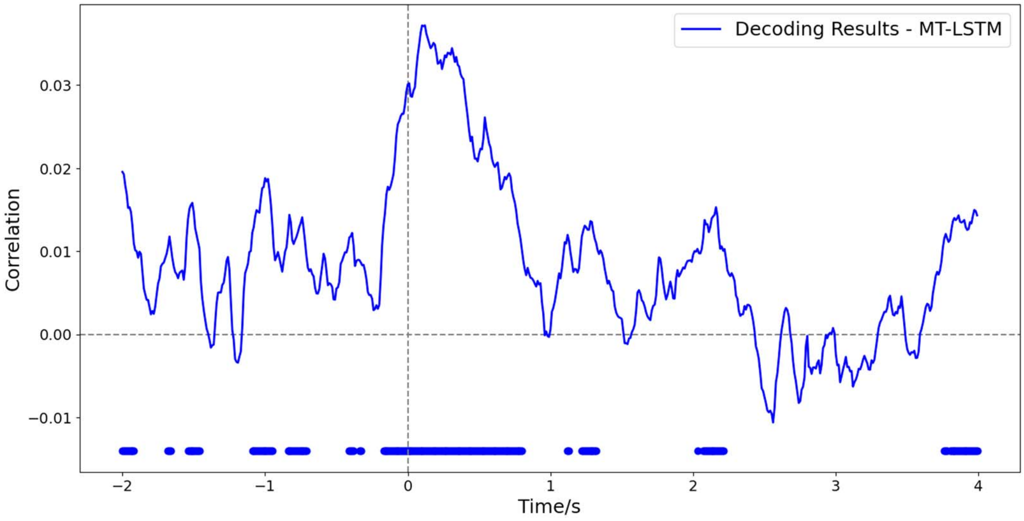


Figure 3
Average correlation between true embeddings and predicted embeddings for MT-LSTM embeddings. Each data point on the line represents the decoding performance of a 0.1 s time window (the point marks the end of the time window). The dots above the x-axis represent significantly above chance predictions ($p < 0.05$, FDR corrected).

4.2 Selecting Subgroups for Different Timescales

In the previous section, prediction performance was derived from the average effect over all MT-LSTM embedding dimensions. However, when considered in aggregate, we cannot distinguish which dimensions contribute most to successful decoding. In the following analysis, we explored the prediction performance for different timescales independently. If the prediction is above chance at one time point for a particular timescale, it may indicate that the information of this timescale is similar to what is being processed in the brain.

Recall that MT-LSTM timescales are determined by the forget gate biases. Because one MT-LSTM dimension carries limited noisy information, we cannot analyze each dimension independently. Therefore, we selected subsets of MT-LSTM neurons and analyzed them as a group. The 1,150 MT-LSTM dimensions have ordered timescales because the forget gate biases are sampled from a fixed range of values according to the inverse gamma distribution, and so neurons with adjacent indices are tuned to similar timescales. We can create contiguous groups of dimensions and then average the prediction performance to extract related properties of the group.

To partition all 1,150 dimensions into groups, we considered the autocorrelation for each dimension of the hidden state. Before a lexical token w_t is input into the MT-LSTM, the hidden state is v_t. After proceeding to the next n tokens, the hidden state becomes v_{t+n}. We calculated the correlation between dimensions of v_t and v_{t+n} across time, namely the *autocorrelation*. A plot of autocorrelation values is given for the dimensions of v and number of words delay n in Figure 4, showing the stability of each dimension. To reduce noise, we average adjacent groups of 5 dimensions. We partitioned MT-LSTM embedding dimensions into 3 groups based on the patterns on autocorrelation plot (Figure 4): The first group (1–8) has a very high (>0.8) autocorrelation ($7K < T < 360K$). The second group (8–400) has a relatively stable autocorrelation that is around 0.6, regardless of number of words delay ($7.4 < T < 7K$). The third group (400–1,150) has an autocorrelation that steadily decreases from around 0.6 to 0 both as the timescale

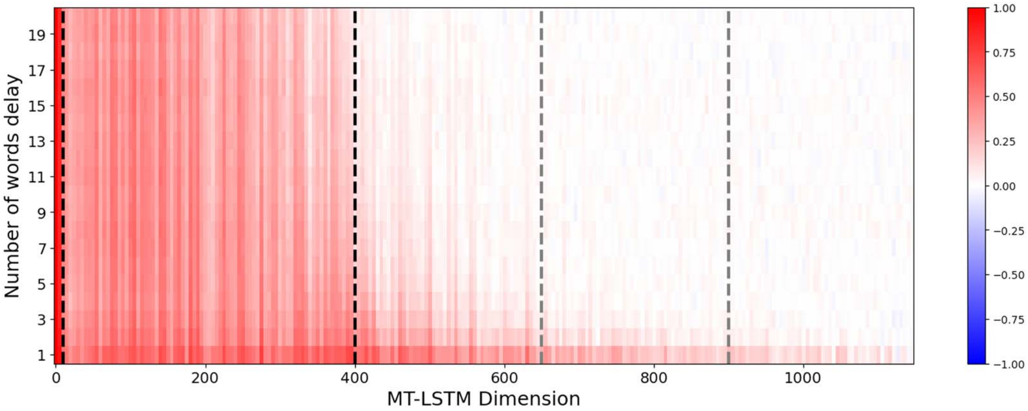


Figure 4
The autocorrelation of all 1,150 MT-LSTM hidden states with delay of up to 20 words. We used this autocorrelation to partition the MT-LSTM dimensions into Long, Medium, and Short timescales (black dashed line), and the Short part further into S-Long, S-Medium, and S-Short timescales (gray dashed line).

decreases and number of words delay increases ($T < 7.4$). We denote these groups: Long, Medium, and Short timescale groups, as shown in Figure 4.

The short timescale group is quite large (750 units), and shows a fairly linear decrease in correlation (also see Figure C.1), implying we should consider its units in a more fine-grained way. We further separated the short timescale into 3 equal subsets as shown in Figure 4. In this partition, the short-long (S-Long) timescales are dimensions 400 to 650 ($2.7 < T < 7.4$). The short-medium (S-Medium) 650 to 900 ($1.2 < T < 2.7$), and the short-short (S-Short) 900 to 1,150 ($T < 1.2$).

4.3 Analysis 2: Decoding Partitioned MT-LSTM Embeddings

Based on the grouping of MT-LSTM embeddings defined in Section 4.2, we can investigate which timescales support good decodability. Figure 5 and Figure C.2 show the average decoding performance with the unbalanced timescale groups. The long timescale group (Figure C.2) has obvious oscillations. The maximum correlation is larger than 0.05 and the minimum is -0.15. However, the results are not significant. For the medium and short timescales (Figure 5), there is no rapid change in prediction performance along the timeline. The medium timescale is only significant in a small period (360 ms) after the onset of the word with a correlation peak of 0.03. The short timescale has the highest peak (0.05) and widest range (1 s) of significant correlation around the onset of word. The significant correlation for the short timescale also spans across the timeline. To conclude, the decreasing decodability for longer timescales goes *against* our hypothesis that longer MT-LSTM timescales are indicative of longer processing time in the brain. We discuss this point further in Section 5.4.

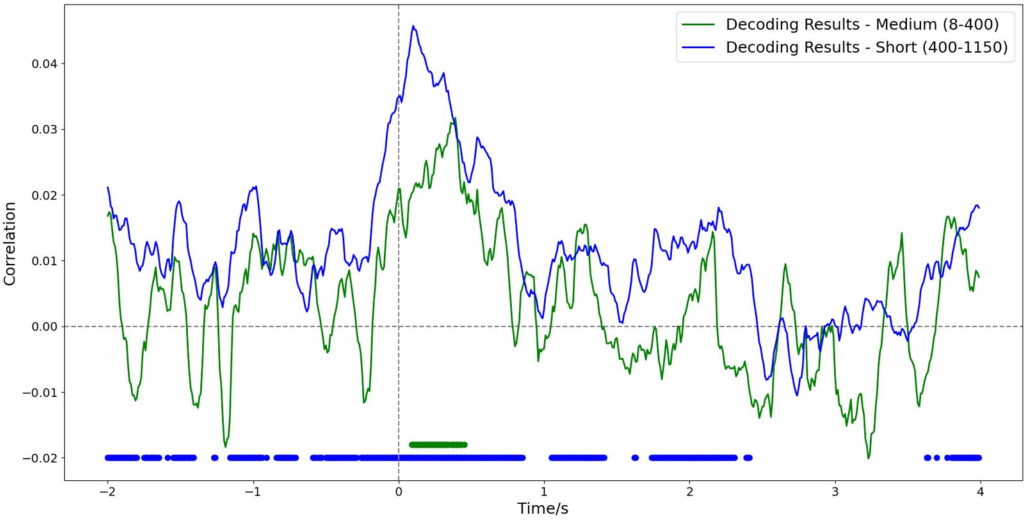


Figure 5
Decoding results based on dimensions extracted from a specific partition of the corresponding MT-LSTM representation. Each data point on the line represents the decoding performance of a 100 ms time window (the point marks the end of the time window). The dots above the x-axis represent significantly better than chance predictions ($p < 0.05$, FDR corrected).

4.4 Analysis 3: Decoding Fine-grained MT-LSTM Embeddings

As described in Section 4.2, to thoroughly explore the short timescale group, we divided it into 3 equally sized subsets (S-Long, S-Medium, and S-Short). The three colored lines in Figure 6 show the corresponding decoding results of those timescales, and the dots along the bottom of the graph indicated above chance decodability. All timescales have a peak around word onset. However, the ranges of above chance predictions differ. The S-Short timescales show significant decodability only around the onset from -180 ms to 790 ms, while significant decodability of the S-Long timescales are widely spread from -2 s to 2 s. The paired cluster permutation test (Maris and Oostenveld 2007) (paired based on the same split of train and test sets during cross validation) contrasting S-Long and S-Short decodability between -1 s and 2 s shows that the 2 conditions are significantly different around the onset ($p = 0.001$). The S-Short timescale decodability has several peaks between 0 s and 1 s, but the S-Long timescale decodability fluctuates less. In general, the timing of the above chance points for each timescale corresponds with the timescale length in the MT-LSTM model.

Although the long timescale group produces the most reliable predictions at distant time points, other timescale groups also produce reliable predictions. For the S-Short timescale, above chance predictions also appear occasionally near $[-1.5, -1.2]$ s. The S-Medium timescale is similar to this pattern, but with more frequent above chance predictions near $[-0.8, -0.5, 1.2]$ s. The S-Long and S-Medium timescale groups also show above chance prediction around 4 s after word onset, i.e. the information in some short timescales reappears in the brain representations. This reappearance may be caused by neighboring words that share mutual information with the center word, but note that the decoding of non-contextual word vectors (Figure B.1) shows no

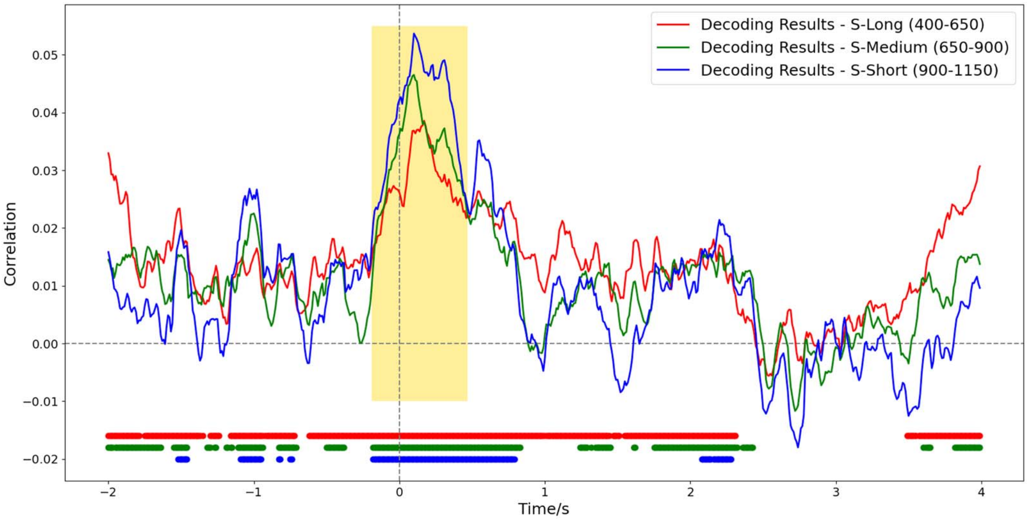


Figure 6
Decoding results for MT-LSTM embeddings based on the partition of short timescales. Each data point on the line represents the decoding performance of a 100 ms time window (the point marks the end of the time window). The dots above the *x*-axis represent significantly better than chance predictions ($p < 0.05$, FDR corrected). The yellow shaded area shows the largest cluster where S-Short has significantly different decodability from S-Long ($p = 0.001$).

above-chance decodability beyond 700 ms after word onset. Thus, this reappearance may also indicate a difference between brain and MT-LSTM representations. We will return to this point further in Sections 5.1–5.3.

4.5 Analysis 4: Temporal Generalization of Decoding Models

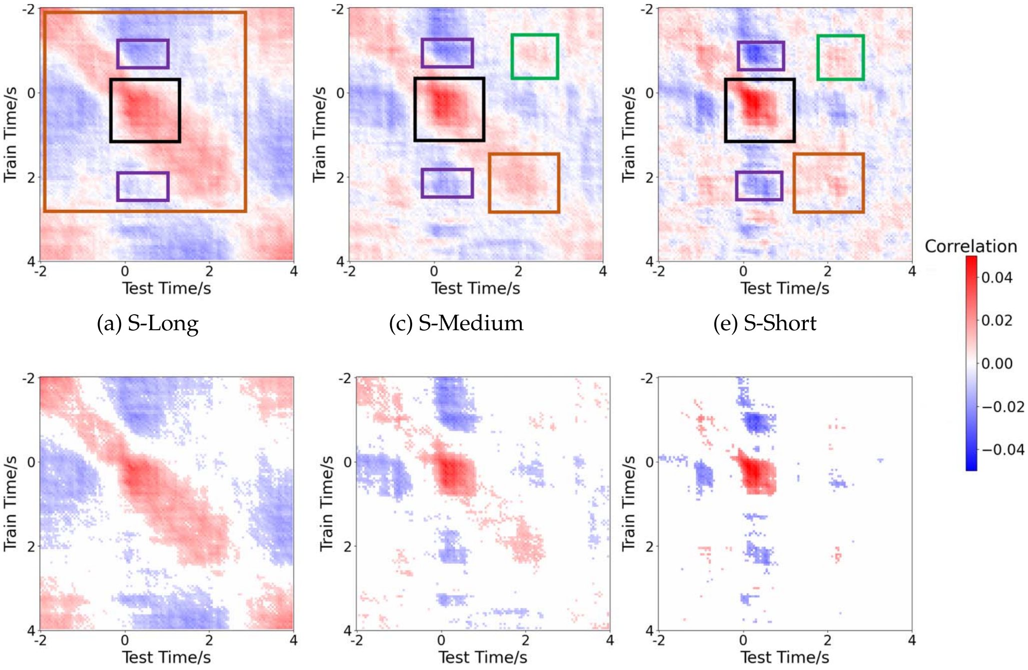
From the timescale analysis in Section 4.4, we see different prediction performance for the S-Long, S-Medium, and S-Short timescales, and that the performance is above chance for different durations. In this section, we use the TGM (Section 3.7) to investigate the generalizability of a model over time. In a TGM analysis, a model trained on one time point is evaluated at other time points. Recall that, if a model trained using one window of EEG data can successfully predict MT-LSTM embeddings using data from another time window, this indicates that the brain representations leveraged by the decoding model are similar. We used this to guide an exploration of each period in Figure 6 where multiple time points were significantly above chance.

We used the same partition of timescales used in Section 4.4, but changed the stride of the sliding window from 10 ms to 50 ms to decrease the computation required to generate the TGMs. The results and the FDR-corrected results for S-Long, S-Medium, and S-Short groups are shown in Figure 7. In order to focus our discussion (and avoid analyzing false positives), we only discuss clusters which are consistent across at least two timescales (i.e., they appear at similar time intersections for at least two of S-Long, S-Medium, and S-Short timescales). Because TGMs are often quite symmetric, for brevity we will refer to only one of the two symmetric clusters, though the effects are usually equivalent for the mirror image.

Figures 7a, 7c, 7e show a similar square-like pattern in the middle (black rectangles, 0 s–1 s on both train and text axes). This indicates that a model trained around word onset generalizes well to similar time points in the following 1 second window. This implies that the brain activity leveraged by the model is similar within this range. The three columns show clear differences in the extent of temporal generalization across time. For the S-Long timescale (Figure 7b), the small square is part of a band-like pattern (brown rectangle, −2 s to 2 s on both train and test axes) that extends from prior to word onset until 2 s after word onset. This symmetric pattern along the diagonal shows that each model on the timeline in this [−2,2] s range can generalize in a 1 second window, except for a brief period near 0 s (possibly due to a change in representation influenced by word perception). These results indicate the persistence of the S-Long timescale representations in the brain.

The TGM for the S-Short timescale (Figure 7e) shows a cluster from [−0.1, 0.75]s (black rectangle) that is separated from nearby clusters. That is, the areas indicated by the black rectangle and brown rectangle (around 2 s on both axes) are not connected, implying that the brain processes S-Short timescales at times close to word onset. The S-Medium timescale (Figure 7c) has a transitional form: There exists not only a square resembling that seen in the short timescale (black rectangle), but also a faint band-like pattern (brown rectangle, from 1 s to 2 s on both axes) reminiscent of the one seen for S-Long.

The TGMs also show reliable predictions for distant time points (points far from the diagonal). For the S-Short timescale (Figure 7e), The models generalize well when training on windows around −1 s and testing on windows around 2 s (green rectangle). However, when we train the models around −1 s or 2 s and test between 0 s and 1 s, the models show a significantly below chance decodability (purple rectangles). For the S-Long timescale (Figure 7a), the regions of below chance decodability are sustained



(a) S-Long

(c) S-Medium

(e) S-Short

(b) S-Long, FDR corrected

(d) S-Medium, FDR corrected

(f) S-Short, FDR corrected

Figure 7
Temporal generalization matrices for S-Long, S-Medium, and S-Short timescales. Top: All correlation values in each TGM. Bottom: Same as above, but only with significant time points colored ($p < 0.05$, FDR-corrected). Black rectangles: The clusters on the main diagonal, where we train and test the decoding model around the word onset (0 s). Brown rectangles: The clusters on the main diagonal, where we train and test the decoding model around 2 s after word onset. Purple rectangles: The clusters where we train the decoding model around 1 s before or 2 s after word onset and test around word onset (0 s). Green rectangles: The clusters where we train the decoding model around 1 s before word onset, and test around 2 s after word onset.

longer with respect to models trained before −1 s and after 2 s, and tested between 0 s and 2 s (purple rectangles). The S-Medium timescale (Figure 7c) is again a transitional form between S-Short and S-long: It is similar to the short timescales but enlarges the generalization window for distant time points (purple and green rectangles).

These results provide us with further evidence that the brain differentially processes long and short timescale information in a way that resembles the MT-LSTM. As the timescale increases, the representations in the brain become more persistent, and TGMs show generalizability across longer time spans. However, the TGMs also show below chance regions, indicating negated representations at distant time points compared to around the onset. We will return to this in Section 5.3.

4.6 Analysis 5: Brain Areas Contributing to Decoding Performance

For this analysis, we trained decoding models using different contiguous sensor subsets to investigate the contribution to decoding performance attributable to spatially grouped brain regions. Figure 8 shows topographic maps of the decoding accuracy at a few key time points which were selected based on reliable decoding accuracy in

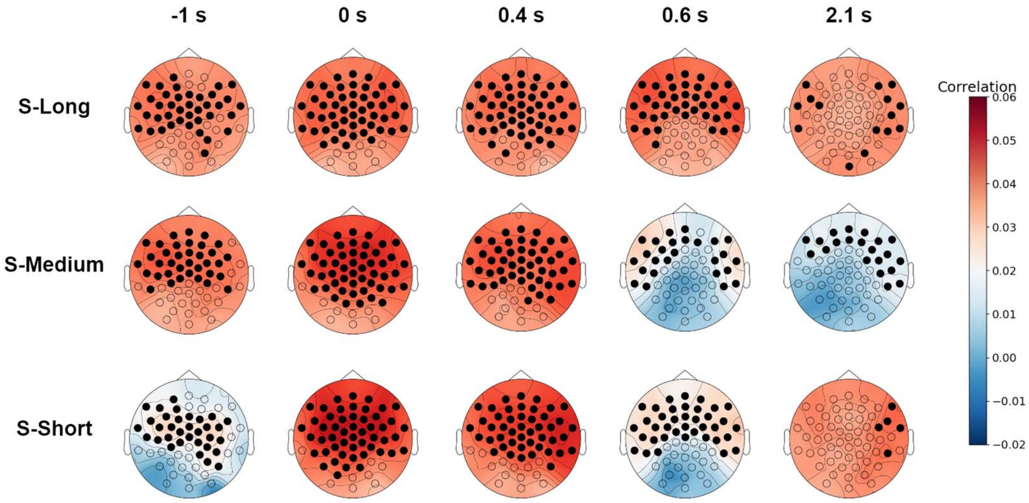


Figure 8
The topographic map at a few key time points for S-Long, S-Medium, and S-Short timescales. All 3 timescales show significant decodability at these time points, according to Figure 6. −1 s and 2.1 s are time points distant from the onset, while 0 s is the onset and 0.4 s and 0.6 s correspond with the time for event-related potentials N400 and P600. Solid circles show significantly better than chance predictions ($p < 0.05$, FDR corrected).

Figure 7. However, due to the low spatial resolution and high noise in EEG, one must proceed with caution. We use these results only to produce some preliminary hypotheses about the differences between long and short timescales.

At 0 s and 0.4 s, for all 3 timescales, we can significantly decode MT-LSTM embeddings at most sensors. The highest accuracy appears towards the frontal lobe from both hemispheres. At 0.6 s, the areas appear towards the temporal lobe and prefrontal cortex. These results correlate with the retrieval-integration cycle in Brouwer and Hoeks (2013), in which the left posterior middle temporal gyrus retrieves word semantics and induces a N400, while the left inferior frontal gyrus integrates sentences and induces a P600. However, our results for [0, 0.4, 0.6] s and all 3 timescales in Figure 8 do not indicate a laterality preference.

At 1 s before word onset (−1 s), for all 3 timescales, we can significantly decode MT-LSTM embeddings at about half the sensors, with the highest accuracy consistently concentrated over the temporal and frontal lobes. Similar to near the onset (0 s), the results for −1 s and all 3 timescales in Figure 8 do not indicate a laterality preference.

At the time window furthest from word onset (2.1 s), we can see that the area for the S-Short timescale is mainly above the right temporal lobe. The S-Medium timescale also shows high decoding performance over the prefrontal cortex. This closely corresponds to findings reported in Jain et al. (2020), in which auditory cortex (temporal lobe) prefers shorter timescales, while the prefrontal cortex corresponds to longer timescales. However, for the S-Long timescale, we did not find significant decodability over prefrontal cortex, instead, the above chance sensors are concentrated over the temporal lobes of both hemispheres. This may be consistent with Jain et al. (2020), in that the temporal lobes have a preference for a wider range of timescales than the prefrontal cortex.

These results show that the brain areas that track long and short timescale information are distinct and correspond with previous findings from earlier research in fMRI

(Jain and Huth 2018). However, due to the low spatial resolution of EEG data, we cannot reliably attribute the source of decodability to more precise brain areas (e.g., the auditory cortex). Studies (Paulesu et al. 1997; Whitney, Jefferies, and Kircher 2011; Hagoort and Indefrey 2014; Hertrich et al. 2021) have also shown that both temporal lobe and prefrontal cortex are functionally heterogeneous and may involve multiple levels of language processing. In addition, decoding at distant time points from word onset (-1 s, 2 s) may also result from decoding neighboring words that share mutual information with the center word. To conclude, these results hint at some possible interpretation for the processing of timescale information in different areas of the brain, but further work will be required to draw more spatially accurate conclusions.

4.7 Results Summary

From our analyses we conclude the following:

1. Information from multiple timescales can be decoded from small windows of EEG (Analysis 1).
2. The Short timescales ($T < 7.4$) can be decoded consistently across the EEG window, while the Medium timescale ($7.4 < T < 7K$) can be decoded only for a short period after word onset (Analysis 2). This motivated us to further analyze the subsets of the short timescale units (S-Long, S-Medium, S-Short).
3. The brain's representation for the S-Long timescale ($2.7 < T < 7.4$) is stable between $[-2, 2]$ seconds, except for a short period near word onset. The representation for the S-Short timescale ($T < 1.2$) is stable only between $[-0.1, 0.75]$ s (Analysis 3, 4).
4. When models trained around distant time points (-1 s and 2 s) are tested around word onset (0 s to 1 s), the models show a significantly below chance decodability, suggesting an inversion of the representation during language comprehension of the timescale information (Analysis 4).
5. Around word onset (0 s and 0.4 s), most sensors can decode S-Long, S-Medium, and S-Short timescale representations. At distant time points (2 s), S-Short timescales can be decoded from the right temporal lobe, while S-Medium timescales can be decoded from the prefrontal cortex (Analysis 5).

5. Discussion

Based on the results described above, we conclude with the following main interpretations:

(1) The duration of decodability after word onset for 2 timescale subgroups, S-Medium and S-Long, is mostly compatible with *forgetting time* in MT-LSTMs. However, the duration of decodability after word onset for S-Short timescale representations persist for longer in the brain than the *forgetting time* in an MT-LSTM would suggest.

(2) Successful decoding of timescale information before word onset shows how mutual information is preserved from context before the onset of upcoming word.

(3) The anti-correlated decoding results at distant time points in the TGM may be explained by oscillations in the brain's neural activity.

(4) We find little evidence for the decodability of Medium time scale MT-LSTM neurons except for directly after word onset. This implies that their role in the MT-LSTM may not be a good parallel to the brain's processing at medium timescales.

5.1 Decodability Windows Correlate with Timescales

Two MT-LSTM timescale sub-groups that we defined in our partition of Short MT-LSTM neurons (S-Medium / S-Long) correspond to time windows of successful decodability that match the timescale T (*forgetting time*)[2] in the MT-LSTM. Considering the speech rate for *Alice* dataset is approximately 3 words per second (Bhattasali et al. 2020), the S-Medium timescale range, $1.2 < T < 2.7$ corresponds with a range of 0.4 s to 0.9 s in the speech stimuli, while the S-Medium timescale range, $2.7 < T < 7.4$ corresponds with a range of 0.9 s to 2.5 s in the speech stimuli. In Sections 4.4 and 4.5, we showed that the S-Medium timescales are predictable using EEG until about 0.8 s after spoken word onset and S-Long Timescales have near continuous decodability until about 2 s after word onset. Therefore, the time lengths for decodability are compatible with the parameter-tuned timescale (*forgetting time*) (Mahto et al. 2020) in the MT-LSTM model.

The S-Short timescale ($T < 1.2$) has the shortest period of continuous decodability after word onset (790 ms). However, from Figure 6, this length of decodability is only slightly shorter than that of the S-Medium timescale, and 790 ms is about twice the forgetting time ($T = 1.2$ corresponds with 0.4 s in speech stimuli). This indicates a difference between the brain and MT-LSTM's embeddings: In MT-LSTMs, the persistence of information is determined by the neuron's corresponding forget gate bias, which forces the rapid decay of information in short timescales. However, the brain appears to retain this information for longer than the length of two words. This extended decoding time window supports the results reported by Sudre et al. (2012) wherein noun decodability was sustained for at least 700 ms (end of their analysis window) after word onset. Fyshe et al. (2019) found that when people read adjective-noun phrases, the representation of the first word was sustained for nearly 1.5 s after the first word's onset, well into and past the presentation of the second word. Our work provides further evidence that the brain retains the information in short timescales longer than the MT-LSTM model.

5.2 Significantly Above-chance Decodability Before Word Onset

In a language model, the ability to predict the next word is based on the principle that words appearing together share mutual information. Similarly, the brain can also predict future words based on mutual information from context. However, previous

[2] In Vo et al. (2023), an LSTM unit with timescale T means it has the ability to remember information over T words.

work on the timing of next word prediction in the brain shows incongruent results: Wlotko and Federmeier (2015) found that when fixing word intervals, it takes at least 250 ms after word onset for predictions to be detectable, though it sometimes happens within a shorter temporal window. Goldstein et al. (2020) argue that when participants are reading a story, neural responses correlating with upcoming word semantics can be detected as early as 800 ms before word onset. Based on our separation of timescales, for S-Long and S-Medium timescales, the mutual information which is critical for prediction appears long before word onset, while for S-short timescales decoding accuracy is most consistently above chance from 180 ms before word onset.

5.3 Inverted Representations Distant from Word Onset

Our TGM analysis in Section 4.5 shows a strong square of above-chance decoding on the diagonal near word onset (black rectangles, Figure 7). However, when the TGM is applied to windows at more distant time points, we observed significantly below-chance results (purple rectangles, Figure 7). This inversion effect shows that the model's performance is significantly *worse* than chance, even when the same model has significantly above chance accuracy when we train and test with the same time window. We speculate that this could be a byproduct of neural circuits producing oscillations as they handle information flow at different timescales (Jensen et al. 2014). For example, in our results an oscillation of around 1 Hz could be related to S-Short timescale information. This frequency cycles once per second and could account for the decoding performance at -1 s, 0 s, 1 s, and 2 s. Ding et al. (2016, 2017); Keitel, Gross, and Kayser (2018); and Kaufeld et al. (2020) discovered neural oscillations of different frequencies related to processing words, phrases, and sentences, which further supports this theory. However, further experiments with different settings (e.g., altering speech rate) will be required to confirm this.

5.4 Limited Decodability of Medium Timescale Neurons

In Section 5.1, we discussed that the time range of successful decoding increases with timescale length. However, in Figure 5 we found only a narrow band of 500 ms post word onset from which we could successfully decode Medium timescales ($7.4 < T < 7K$). Mahto et al. (2020) found that ablating neurons of these timescales decreases the MT-LSTM's performance in predicting the low-frequency words. Thus, these neurons may store information related to low-frequency words. Chien et al. (2020) also found some "integrator" neurons in their LSTM. Ablating these "integrator" neurons only reduces the prediction accuracy for the last words in a sentence. These results indicate that these neurons with special functions are less frequently activated than other medium-timescale neurons (e.g., the "controller" neurons in Chien et al. (2020), which reduce overall prediction accuracy when ablated. If the intervals between decodability of the medium timescale neurons are longer than our test window (6 s, about 18 words), we cannot detect above chance decodability far from word onset. Therefore, for medium timescales, further experiments may be required to more carefully explore the decodability of subsets of medium timescale neurons further from word onset.

5.5 Limitations

In the decoding analysis described in Section 3.3, we discarded function words and did not report decoding accuracy for them. Function words do carry semantic information, however, they are frequent, and thus LSTM predictions for them are relatively easy and often do not require extended context. Conversely, predicting content (non-function) words requires the LSTM model to integrate a broader range of context. To include function words would skew the results towards shorter contexts.

In Section 4.3, we partitioned the 1,150 MT-LSTM units unevenly into Long, Medium and Short timescale groups with 8, 392, and 750 units. The Short timescale group has the widest range of significant decodability across the timeline, while the Medium timescale only showed decodability for 360 ms, and we were not able to decode information in Long timescales. A possible concern is that our results seem to imply that there is *no* corresponding Long timescale of information in the brain. However, in our partition of timescales, even the shortest of the long timescale MT-LSTM neurons has $T > 7K$, a timescale that is longer than our stimulus story (2.1K words). Thus, our experiment was likely not the right setting to detect such long timescales, and future work will be needed to explore longer stories.

Another concern related to the partitioning of timescales is that the range of significant decodability may be an artifact of unequal statistical power: a group with more samples (MT-LSTM units) might result in better decodability. In Section 4.4, the S-Long and S-Medium groups both have 250 units (fewer than 392 in Medium group), but they have a wider range of significant decodability than the Medium group. This is evidence that the reduced decodability of the Medium group is not an artifact of reduced sample size compared to Short group. To further investigate this, we partitioned the Long and Medium group into 2 parts, 0–250, and 150–400, which have the same sample size (250) as the S-Long (400–650) group. The decoding results in Figure C.3 show that equalizing the number of units does not change the decodability: The 0–250 group shows significant decodability nowhere on the timeline (resembling Long timescale group), and the 150–400 group shows significant decodability mainly around onset (resembling Medium timescale group and having a narrower range than S-Long group). These results all indicate that the range of significant decodability in Medium timescales is likely not an artifact of the number of units.

In Section 4.4, the correlation values between the MT-LSTM embeddings and EEG-decoded embeddings are very low. This could be due to multiple factors. The first factor is the low SNR of EEG. In the dataset, the story was played only once to each participant, so each word in the specific context was presented one time to each participant, and we cannot average across words at analysis time. This negatively impacts the SNR in the EEG data supplied to the decoding model. Another possible reason is that the participants are listening passively to the story and the language model is simply predicting the next token. Oota et al. (2022) show that the brain-model correspondence is influenced by human and model tasks. Finally, there are also studies by Caucheteux and King (2022) and Antonello and Huth (2024) who suggest there are significant discrepancies between brains and models in the domain of language. These factors likely all contribute to the lower decoding results, but it is difficult to determine the extent to which each factor is responsible.

6. Conclusion

In this article, we explored the temporal sensitivity of language processing in the brain using an MT-LSTM, a neural network model of language processing that explicitly encodes varying timescales. Leveraging EEG's high temporal precision, we examined the relationship between short, medium, and long MT-LSTM representations and tested when dimensions of these timescales could be decoded using windows of EEG around word onset. We explored the generalizability of these decoding models and found evidence for both stability and negation of brain representations during the processing of language. These preliminary results with high temporal resolution EEG data add to an existing body of research, which has until now primarily used fMRI to study the representation of linguistic timescales in the brain. Our results may inform the design of future experiments to understand how multiple levels of temporally sensitive language processing are represented in the brain.

Appendix A. Preprocessing EEG

Forty-nine native English speakers (14 male, aged 18–29 years) participated in the study. Eight participants were excluded due to excessive noise. A further 8 were excluded due to low comprehension questionnaire performance and 4 participants were excluded due to data corruption issues during EEG acquisition. For each participant, we used the following preprocessing method to remove noise and artifacts using the MNE toolbox (Gramfort et al. 2013):

1. Re-referencing to the average of both mastoid electrodes.
2. Applied a [0.1, 100] Hz FIR band filter and a 60 Hz notch filter to remove line noise.
3. Manual identification of bad channels (high impedance or noisy segments).
4. Used Spherical Spline Interpolation (SSI) to repair bad channels.
5. Applied Independent Component Analysis (ICA) to identify and remove ocular as well as other physiological artifacts (cardiac, muscular etc.)

As described in Section 3.3, we chose 800 tokens from 914 lexical words. For each participant and each token, a -2 s to 4 s time window around word onset was selected. Time windows were rejected if the maximum amplitude in this window was larger than 80 µV. For each token, the EEG was the average (across participants) of all remaining instances of this token that were not rejected.

During this process, we found that the number of good segments for some participants ($n = 6$) was smaller than 300. For some participants ($n = 4$), after averaging all the good segments, average EEG signals deviated from baseline or showed obvious alpha waves. These phenomena indicated the inattention of the participants or the unsatisfactory quality of the data. Therefore, we also excluded these participants whose data exhibited these effects.

Finally, each token used in our analysis consists of a 59 channel × 6 second time window of EEG, averaged over all good segments for all 19 participants. A linear detrending step is applied on every channel in the 6 second window in order to derive the data matrix used in our analysis.

Appendix B. Decoding Non-contextual Word Vectors

The EEG dataset was collected continuously while participants were presented with uninterrupted speech. Recorded EEG activity captures a broad range of activity in the

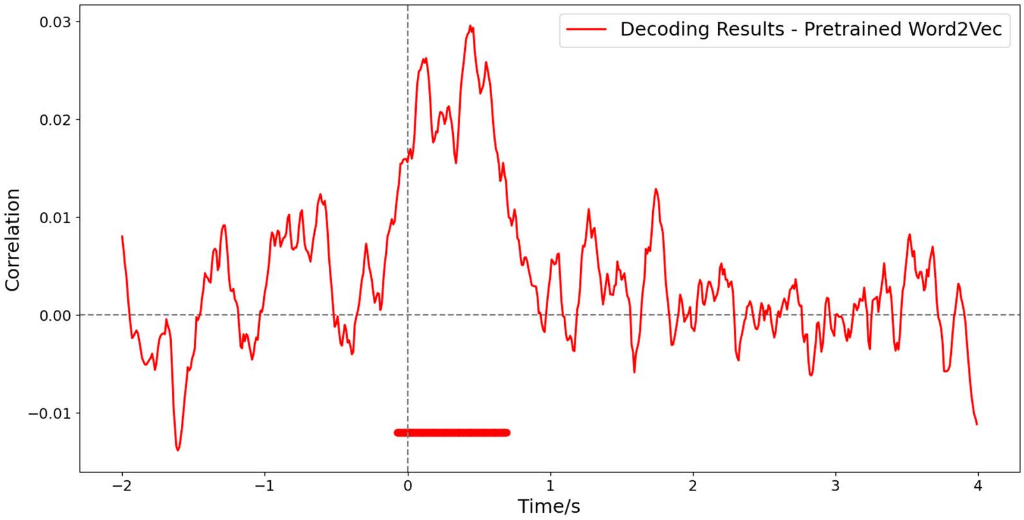


Figure B.1
Average correlation between true embeddings and predicted embeddings for pretrained Word2Vec vectors. Each data point on the line represents the decoding performance of a 0.1 s time window (the point marks the end of the time window). The dots above the *x*-axis represent significantly better than chance predictions ($p < 0.05$, FDR corrected).

brain, beyond specifically language processing. We performed a baseline analysis to ensure the preprocessed EEG data contain decodable *non-contextual* semantic information before we formally compare them with *contextual* embeddings tuned to different timescales. We chose the 300-dimension pretrained Word2Vec vectors trained on the Google News dataset using the Skipgram algorithm (Mikolov et al. 2013). Analyses are identical to that described above for contextual embeddings. If decoding performance is significantly better than chance around word onset, we can conclude that the preprocessed EEG signals carry semantic information and it is appropriate to perform a decoding analysis with timescale-tuned representations.

The result for decoding non-contextual word embeddings is shown in Figure 3. We found that reliable predictions appear from 70 ms before the onset and last until 690 ms after the onset. The curve has multiple peaks at 200 ms and 500 ms. This result shows that non-contextual semantic information can be decoded from EEG signals around word onset. Considering the average duration of a selected lexical word is about 391 ms in *Alice* dataset, the period lasting to 690 ms indicate the brain's further processing after the offset of the spoken word.

Appendix C. Supplementary Figures

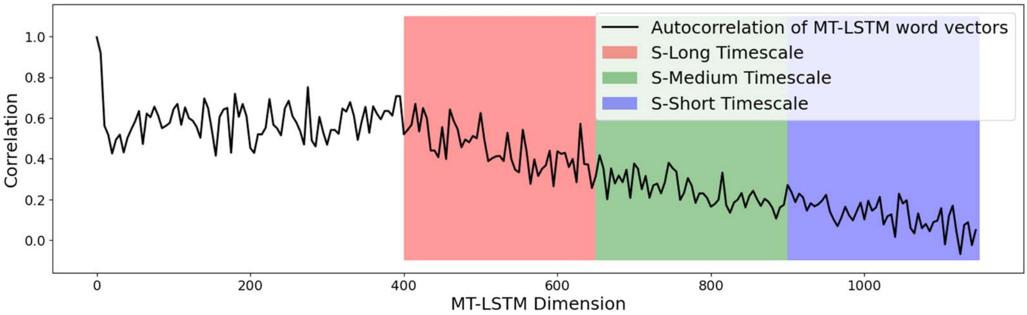


Figure C.1
Autocorrelation of MT-LSTM hidden states with delay of 1 word and our partitioning of Short timescales into S-Long, S-Medium, and S-Short timescales.

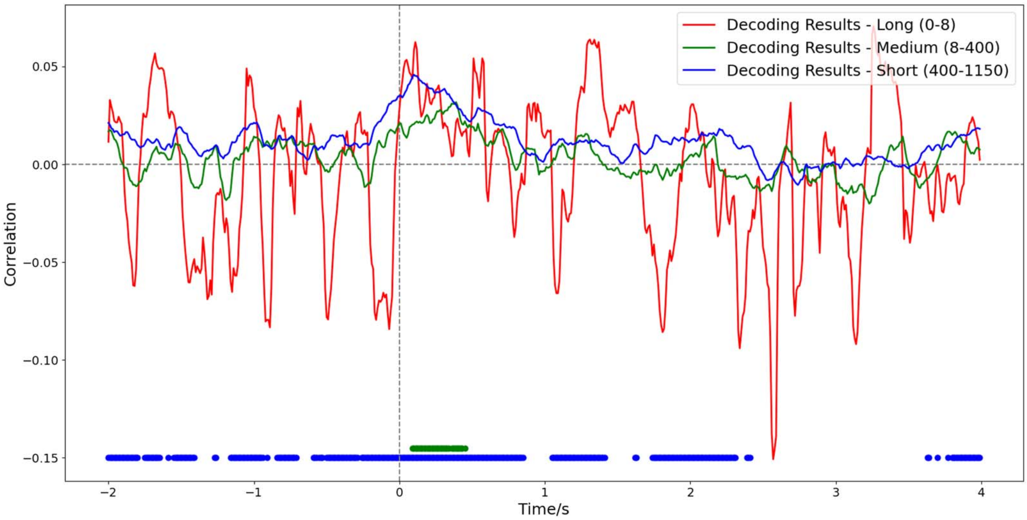


Figure C.2
Decoding results for MT-LSTM embeddings based on the partition: Long (0–8), Medium (8–400), and Short (400–1,150). Each data point on the line represents the decoding performance of a 100 ms time window (the point marks the end of the time window). The dots above the *x*-axis represent significantly better than chance predictions ($p < 0.05$, FDR corrected).

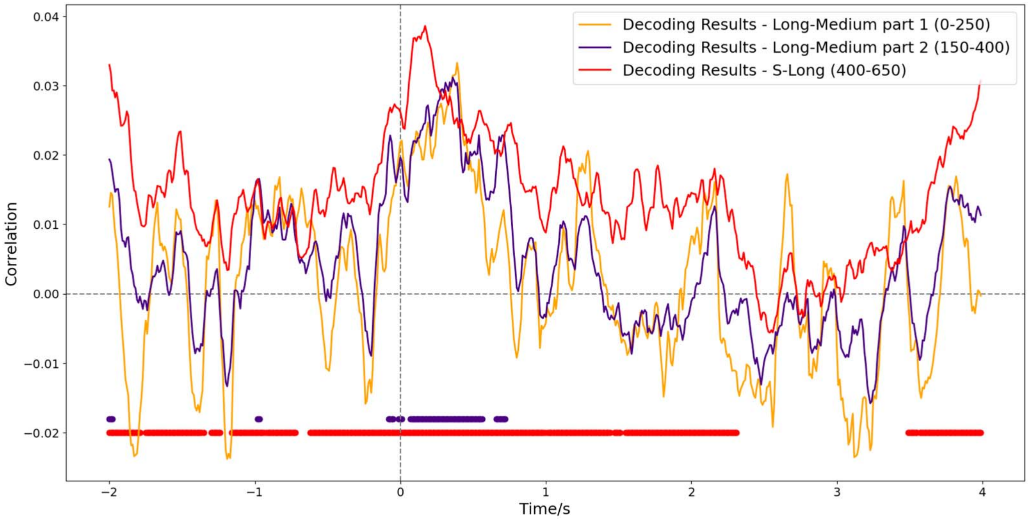


Figure C.3
Decoding results for MT-LSTM embeddings based on a new partition for Long and Medium timescales that ensures equal sample size (250): 0–250, 150–400, and 400–650 (The original S-Long group). Each data point on the line represents the decoding performance of a 100 ms time window (the point marks the end of the time window). The dots above the x-axis represent significantly better than chance predictions ($p < 0.05$, FDR corrected).

Acknowledgments

This research was supported by Alberta Machine Intelligence Institute (AMII), the Canadian Institute for Advanced Research (CIFAR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Digital Research Alliance of Canada (alliancecan.ca). Alona Fyshe holds a Canada CIFAR AI Chair. We express our gratitude to the editors and anonymous reviewers for their helpful comments and suggestions.

References

Amari, Shun-ichi. 1993. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4–5):185–196. https://doi.org/10.1016/0925-2312(93)90006-O

Antonello, Richard and Alexander Huth. 2024. Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1):64–79. https://doi.org/10.1162/nol_a_00087, PubMed: 38645616

Benjamini, Yoav and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bhattasali, Shohini, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. 2020. The Alice Datasets: fMRI & EEG observations of natural language comprehension. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 120–125.

Blank, Idan A. and Evelina Fedorenko. 2020. No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, 219:116925. https://doi.org/10.1016/j.neuroimage.2020.116925, PubMed: 32407994

Brennan, Jonathan and Liina Pylkkänen. 2012. The time-course and spatial distribution of brain activity associated with sentence processing. *NeuroImage*, 60(2):1139–1148. https://doi.org/10.1016/j.neuroimage.2012.01.030, PubMed: 22248581

Brouwer, Harm and John C. J. Hoeks. 2013. A time and place for language comprehension: Mapping the n400 and the p600 to a minimal cortical network. *Frontiers in Human Neuroscience*, 7:758.

https://doi.org/10.3389/fnhum.2013.00758

Caucheteux, Charlotte, Alexandre Gramfort, and Jean-Remi King. 2021. Disentangling syntax and semantics in the brain with deep networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1336–1348.

Caucheteux, Charlotte and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134. https://doi.org/10.1038/s42003-022-03036-1, PubMed: 35173264

Chen, Catherine, Tom Dupré la Tour, Jack Gallant, Dan Klein, and Fatma Deniz. 2024. The cortical representation of language timescales is shared between reading and listening. *Communications Biology*, 7:Article 284. https://doi.org/10.1038/s42003-024-05909-z, PubMed: 38454134

Chen, Yen Chi. 2017. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187. https://doi.org/10.1080/24709360.2017.1396742

Chien, Hsiang Yun Sherry and Christopher J. Honey. 2020. Constructing and forgetting temporal context in the human cerebral cortex. *Neuron*, 106(4):675–686. https://doi.org/10.1016/j.neuron.2020.02.013, PubMed: 32164874

Chien, Hsiang Yun Sherry, Jinhan Zhang, Christopher Honey, et al. 2020. Mapping the timescale organization of neural language models. *arXiv preprint arXiv:2012.06717*.

Dehaene, Stanislas and Jean-Remi King. 2016. Decoding the dynamics of conscious perception: The temporal generalization method. In *Micro-, Meso- and Macro-dynamics of the Brain*. Springer, Cham, pages 85–97. https://doi.org/10.1007/978-3-319-28802-4_7, PubMed: 28590686

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ding, Nai, Lucia Melloni, Aotian Yang, Yu Wang, Wen Zhang, and David Poeppel. 2017. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in Human Neuroscience*, 11:481. https://doi.org/10.3389/fnhum.2017.00481, PubMed: 29033809

Ding, Nai, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1):158–164. https://doi.org/10.1038/nn.4186, PubMed: 26642090

Farbood, Morwaread M., David J. Heeger, Gary Marcus, Uri Hasson, and Yulia Lerner. 2015. The neural processing of hierarchical structure in music and speech at different timescales. *Frontiers in Neuroscience*, 9:157. https://doi.org/10.3389/fnins.2015.00157, PubMed: 26029037

Fyshe, Alona. 2020. Studying language in context using the temporal generalization method. *Philosophical Transactions of the Royal Society B*, 375(1791):20180531. https://doi.org/10.1098/rstb.2018.0531, PubMed: 31840577

Fyshe, Alona, Gustavo Sudre, Leila Wehbe, Nicole Rafidi, and Tom M. Mitchell. 2019. The lexical semantics of adjective–noun phrases in the human brain. *Human Brain Mapping*, 40(15):4457–4469. https://doi.org/10.1002/hbm.24714, PubMed: 31313467

Gao, Richard, Ruud L. van den Brink, Thomas Pfeffer, and Bradley Voytek. 2020. Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture. *eLife*, 9:e61277. https://doi.org/10.7554/eLife.61277, PubMed: 33226336

Goldstein, Ariel, Eric Ham, Samuel A. Nastase, Zaid Zada, Avigail Grinstein-Dabus, Bobbi Aubrey, Mariano Schain, Harshvardhan Gazula, Amir Feder, Werner Doyle, et al. 2022a. Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. *BioRxiv*. https://doi.org/10.1101/2022.07.11.499562

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2020. Thinking ahead: Spontaneous prediction in context as a keystone of language in humans and machines. *BioRxiv*. https://doi.org/10.1101/2020.12.02.403477

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022b. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380. https://doi.org/10.1038/s41593-022-01026-4, PubMed: 35260860

Gramfort, Alexandre, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. 2013. MEG and EEG data analysis with MNE-python. *Frontiers in Neuroscience*, page 267. https://doi.org/10.3389/fnins.2013.00267, PubMed: 24431986

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. https://doi.org/10.18653/v1/N18-1108

Hagoort, Peter and Peter Indefrey. 2014. The neurobiology of language beyond single words. *Annual Review of Neuroscience*, 37:347–362. https://doi.org/10.1146/annurev-neuro-071013-013847, PubMed: 24905595

Hale, John, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736. https://doi.org/10.18653/v1/P18-1254

Hasson, Uri, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. 2008. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008, PubMed: 18322098

Heilbron, Micha, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P. De Lange. 2022. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119. https://doi.org/10.1073/pnas.2201968119, PubMed: 35921434

Hertrich, Ingo, Susanne Dietrich, Corinna Blum, and Hermann Ackermann. 2021. The role of the dorsolateral prefrontal cortex for speech and language processing. *Frontiers in Human Neuroscience*, 15:645209. https://doi.org/10.3389/fnhum.2021.645209, PubMed: 34079444

Hoerl, Arthur E. and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634

Honey, Christopher J., Thomas Thesen, Tobias H. Donner, Lauren J. Silbert, Chad E. Carlson, Orrin Devinsky, Werner K. Doyle, Nava Rubin, David J. Heeger, and Uri Hasson. 2012. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, 76(2):423–434. https://doi.org/10.1016/j.neuron.2012.08.011, PubMed: 23083743

Honnibal, M. and I. Montani. 2017. Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *Unpublished software application*. https://spacy.io

Huth, Alexander G., Wendy A. De Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458. https://doi.org/10.1038/nature17637, PubMed: 27121839

Hwang, Kyuyeon and Wonyong Sung. 2017. Character-level language modeling with hierarchical recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5720–5724. https://doi.org/10.1109/ICASSP.2017.7953252

Jain, Shailee and Alexander Huth. 2018. Incorporating context into language encoding models for fMRI. *Advances in Neural Information Processing Systems*, 31:6628–6637. https://doi.org/10.1101/327601

Jain, Shailee, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S. Turek, and Alexander Huth. 2020. Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. In *Advances in Neural Information Processing Systems*, volume 33, pages 13738–13749. https://doi.org/10.1101/2020.10.02.324392

Jensen, Ole, Bart Gips, Til Ole Bergmann, and Mathilde Bonnefond. 2014. Temporal coding organized by coupled alpha and gamma oscillations prioritize visual processing. *Trends in Neurosciences*, 37(7):357–369. https://doi.org/10.1016/j.tins.2014.04.001, PubMed: 24836381

Kaufeld, Greta, Hans Rutger Bosker, Sanne Ten Oever, Phillip M. Alday, Antje S. Meyer, and Andrea E. Martin. 2020. Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *Journal of Neuroscience*, 40(49):9467–9475. https://doi.org/10.1523/JNEUROSCI.0302-20.2020, PubMed: 33097640

Kazanina, Nina and Alessandro Tavano. 2023. What neural oscillations can and cannot do for syntactic structure building. *Nature Reviews Neuroscience*, 24(2):113–128. https://doi.org/10.1038/s41583-022-00659-5, PubMed: 36460920

Keitel, Anne, Joachim Gross, and Christoph Kayser. 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, 16(3):e2004473. https://doi.org/10.1371/journal.pbio.2004473, PubMed: 29529019

Kementchedjhieva, Yova and Adam Lopez. 2018. 'Indicatements' that character language models learn English morpho-syntactic units and regularities. *arXiv preprint arXiv:1809.00066*. https://doi.org/10.18653/v1/W18-5417

Kriegeskorte, Nikolaus and Rogier A. Kievit. 2013. Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412. https://doi.org/10.1016/j.tics.2013.06.007, PubMed: 23876494

Lakretz, Yair, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. *arXiv preprint arXiv:1903.07435*. https://doi.org/10.18653/v1/N19-1002

Lerner, Yulia, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011, PubMed: 21414912

Lin, Henry W. and Max Tegmark. 2016. Critical behavior from deep dynamics: A hidden dimension in natural language. *arXiv preprint arXiv:1606.06737*.

Lin, Rui, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. 2015. Hierarchical recurrent neural network for document modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 899–907. https://doi.org/10.18653/v1/D15-1106

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. https://doi.org/10.1162/tacl_a_00115

Mahto, Shivangi, Vy A. Vo, Javier S. Turek, and Alexander G. Huth. 2020. Multi-timescale representation learning in LSTM language models. *arXiv preprint arXiv:2009.12727*.

Maris, Eric and Robert Oostenveld. 2007. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1):177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024, PubMed: 17517438

Merity, Stephen, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Point sentinel mixture models. In *International Conference on Learning Representations*, pages 1851–1865.

Meyer, Lars. 2018. The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience*, 48(7):2609–2621. https://doi.org/10.1111/ejn.13748, PubMed: 29055058

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119.

Mitchell, Tom M., Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195. https://doi.org/10.1126/science.1152876, PubMed: 18511683

Murphy, Alex, Bernd Bohnet, Ryan McDonald, and Uta Noppeney. 2022. Decoding part-of-speech from human EEG signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2201–2210. https://doi.org/10.18653/v1/2022.acl-long.156, PubMed: 35258805

Murray, John D., Alberto Bernacchia, David J. Freedman, Ranulfo Romo, Jonathan D. Wallis, Xinying Cai, Camillo Padoa-Schioppa, Tatiana Pasternak, Hyojung Seo, Daeyeol Lee, et al. 2014. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, 17(12):1661–1663. https://doi.org/10.1038/nn.3862, PubMed: 25383900

ten Oever, Sanne, Sara Carta, Greta Kaufeld, and Andrea E. Martin. 2022. Neural tracking of phrases in spoken language comprehension is automatic and task-dependent. *eLife*, 11:e77468. https://doi.org/10.7554/eLife.77468, PubMed: 35833919

Oota, Subba Reddy, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Raju Surampudi. 2022. Neural language taskonomy: Which NLP tasks are the most predictive of fMRI brain activity? *arXiv preprint arXiv:2205.01404*. https://doi.org/10.18653/v1/2022.naacl-main.235

Paulesu, Eraldo, Ben Goldacre, Paola Scifo, Stefano F. Cappa, Maria Carla Gilardi, Isabella Castiglioni, Daniela Perani, and Ferruccio Fazio. 1997. Functional heterogeneity of left inferior frontal cortex as revealed by fMRI. *Neuroreport*, 8(8):2011–2016. https://doi.org/10.1097/00001756-199705260-00042, PubMed: 9223094

Pereira, Francisco, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963. https://doi.org/10.1038/s41467-018-03068-4, PubMed: 29511192

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

Rafidi, Nicole S. 2018. *Using Machine Learning for Time Series to Elucidate Sentence Processing in the Brain*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Raut, Ryan V., Abraham Z. Snyder, and Marcus E. Raichle. 2020. Hierarchical dynamics as a macroscopic organizing principle of the human brain. *Proceedings of the National Academy of Sciences*, 117(34):20890–20897. https://doi.org/10.1073/pnas.2003383117, PubMed: 32817467

Reddy, Aniketh Janardhan and Leila Wehbe. 2021. Can fMRI reveal the representation of syntactic structure in the brain? In *Advances in Neural Information Processing Systems*, volume 34, pages 9843–9856. https://doi.org/10.1101/2020.06.16.155499

Shen, Yikang, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2018. Ordered neurons: Integrating tree structures into recurrent neural networks. *arXiv preprint arXiv:1810.09536*.

Singh, Moirangthem Dennis and Minho Lee. 2017. Temporal hierarchies in multilayer gated recurrent neural networks for language models. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2152–2157. https://doi.org/10.1109/IJCNN.2017.7966115

Spitmaan, Mehran, Hyojung Seo, Daeyeol Lee, and Alireza Soltani. 2020. Multiple timescales of neural dynamics and integration of task-relevant signals across cortex. *Proceedings of the National Academy of Sciences*, 117(36):22522–22531. https://doi.org/10.1073/pnas.2005993117, PubMed: 32839338

Sudre, Gustavo, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1):451–463. https://doi.org/10.1016/j.neuroimage.2012.04.048, PubMed: 22565201

Tallec, Corentin and Yann Ollivier. 2018. Can recurrent neural networks warp time? *arXiv preprint arXiv:1804.11188*.

la Tour, Tom Dupré, Michael Eickenberg, Anwar O. Nunez-Elizalde, and Jack L. Gallant. 2022. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728. https://doi.org/10.1016/j.neuroimage.2022.119728, PubMed: 36334814

Traxler, Matthew J. 2011. *Introduction to Psycholinguistics: Understanding Language Science*. John Wiley & Sons.

Vo, Vy Ai, Shailee Jain, Nicole Beckage, Hsiang-Yun Sherry Chien, Chiadika Obinwa, and Alexander G. Huth. 2023. A unifying computational account of temporal context effects in language across the human cortex. *bioRxiv*. https://doi.org/10.1101/2023.08.03.551886

Wehbe, Leila, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS ONE*, 9(11):e112575. https://doi.org/10.1371/journal.pone.0112575, PubMed: 25426840

Wehbe, Leila, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014b. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243. https://doi.org/10.3115/v1/D14-1030

Whitney, Carin, Elizabeth Jefferies, and Tilo Kircher. 2011. Heterogeneity of the left temporal lobe in semantic representation and control: Priming multiple versus single meanings of ambiguous words. *Cerebral Cortex*, 21(4):831–844. https://doi.org/10.1093/cercor/bhq148, PubMed: 20732899

Wlotko, Edward W. and Kara D. Federmeier. 2015. Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*, 68:20–32. https://doi.org/10.1016/j.cortex.2015.03.014, PubMed: 25987437

Xu, Jiang, Stefan Kemeny, Grace Park, Carol Frattali, and Allen Braun. 2005. Language in context: Emergent features of word, sentence, and narrative comprehension. *NeuroImage*, 25(3):1002–1015. https://doi.org/10.1016/j.neuroimage.2004.12.013, PubMed: 15809000

Zeraati, Roxana, Yan-Liang Shi, Nicholas A. Steinmetz, Marc A. Gieselmann, Alexander Thiele, Tirin Moore, Anna Levina, and Tatiana A. Engel. 2023. Intrinsic timescales in the visual cortex change with selective attention and reflect spatial connectivity. *Nature Communications*, 14(1):1858. https://doi.org/10.1038/s41467-023-37613-7, PubMed: 37012299