Entropy- and Distance-Based Predictors From GPT-2 Attention Patterns Predict Reading Times Over and Above GPT-2 Surprisal

Byung-Doh Oh
Department of Linguistics
The Ohio State University
oh.531@osu.edu

William Schuler
Department of Linguistics
The Ohio State University
schuler.77@osu.edu

Abstract

Transformer-based large language models are trained to make predictions about the next word by aggregating representations of previous tokens through their self-attention mechanism. In the field of cognitive modeling, such attention patterns have recently been interpreted as embodying the process of cue-based retrieval, in which attention over multiple targets is taken to generate interference and latency during retrieval. Under this framework, this work first defines an entropy-based predictor that quantifies the diffuseness of self-attention, as well as distance-based predictors that capture the incremental change in attention patterns across timesteps. Moreover, following recent studies that question the informativeness of attention weights, we also experiment with alternative methods for incorporating vector norms into attention weights. Regression experiments using predictors calculated from the GPT-2 language model show that these predictors deliver a substantially better fit to held-out self-paced reading and eye-tracking data over a rigorous baseline including GPT-2 surprisal. Additionally, the distance-based predictors generally demonstrated higher predictive power, with effect sizes of up to 6.59 ms per standard deviation on self-paced reading times (compared to 2.82 ms for surprisal) and 1.05 ms per standard deviation on eye-gaze durations (compared to 3.81 ms for surprisal).

1 Introduction

Much work in broad-coverage sentence processing has focused on studying the role of expectation operationalized in the form of surprisal (Hale, 2001; Levy, 2008) using language models (LMs) to define a conditional probability distribution of a word given its context (Smith and Levy, 2013; Goodkind and Bicknell, 2018). Recently, this has included Transformer-based language models (Wilcox et al., 2020; Merkx and Frank, 2021; Oh et al., 2022).

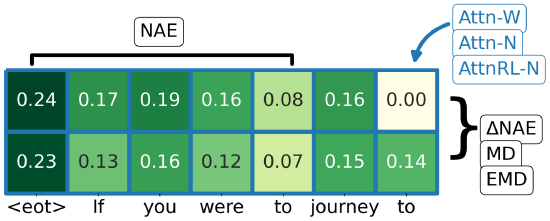

Figure 1: The predictors and attention weight formulations examined in this work. The entropy-based predictor (i.e. NAE) quantifies the diffuseness of attention over previous tokens at a given timestep, while the distance-based predictors (i.e. ΔNAE, MD, EMD) capture the change in attention patterns across consecutive timesteps (top row: weights at 'journey,' bottom row: weights at 'to'). These predictors can be calculated from attention weights formulated using different methods (i.e. ATTN-W, ATTN-N, ATTNRL-N).

However, expectation-based accounts have empirical shortcomings, such as being unable to fully account for garden-path effects (van Schijndel and Linzen, 2021) or predict the timing of delays in certain constructions (Levy et al., 2013). For this reason, some research has begun to focus on the effects of memory and attention using predictors calculated from language model representations. For example, Ryu and Lewis (2021) recently drew connections between the self-attention patterns of Transformers (Vaswani et al., 2017) and cue-based retrieval models of sentence comprehension (e.g. Lewis et al., 2006). Their attention entropy, which quantifies the diffuseness of the attention weights over previous tokens, showed patterns that are consistent with similarity-based interference observed during the processing of subject-verb agreement. However, these results relied on identifying one attention head specialized for the *nsubj* dependency, and an aggregated version of this predictor was not very strong in predicting naturalistic reading times in the presence of a surprisal predictor (Ryu and Lewis, 2022).

This work therefore defines and evaluates several

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9324 - 9334
December 7-11, 2022

memory- and attention-based predictors derived from the self-attention patterns of the Transformer-based GPT-2 language model (Radford et al., 2019) on two naturalistic datasets, in the presence of a strong GPT-2 surprisal baseline. First, normalized attention entropy expands upon Ryu and Lewis's (2021) attention entropy by re-normalizing the attention weights and controlling for the number of tokens in the previous context. Additionally, three distance-based predictors that quantify the shift in attention patterns across consecutive timesteps are presented, based on the idea that the reallocation of attentional focus entails processing difficulty.

Moreover, motivated by work on interpreting large language models that question the connection between attention weights and model predictions (e.g. Jain and Wallace, 2019), the norm-based analysis of the transformed vectors (Kobayashi et al., 2020, 2021) is newly applied to GPT-2 in this work to inform alternative formulations of attention weights. For example, it has been observed that while large language models tend to place high attention weights on special tokens (e.g. <|endoftext|> of GPT-2), these tokens contribute very little to final model predictions as their 'value' vectors have near-zero norms (Kobayashi et al., 2020). Attention weight formulations that incorporate the norms of the transformed vectors should therefore alleviate the over-representation of such special tokens and represent the contribution of each token more accurately.

Results from regression analyses using these predictors show significant and substantial effects in predicting self-paced reading times and eye-gaze durations during naturalistic reading, even in the presence of a robust surprisal predictor. Additionally, alternative formulations of attention weights that incorporate the norms of the transformed vectors are shown to further improve the predictive power of these predictors.

2 Background

This section provides a mathematical definition of the self-attention mechanism underlying the GPT-2 language model and describes alternative norm-based formulations of attention weights.

2.1 Masked Self-Attention of GPT-2 Language Models

GPT-2 language models (Radford et al., 2019) use a variant of a multi-layer Transformer decoder proposed in Vaswani et al. (2017). Each decoder layer consists of a masked self-attention block and a feed-forward neural network:

$$\mathbf{x}_{l+1,i} = \mathrm{FF}(\mathrm{LN}_{\mathrm{out}}(\mathbf{o}_{l,i}+\mathbf{x}_{l,i}))+(\mathbf{o}_{l,i}+\mathbf{x}_{l,i}), \quad (1)$$

where $\mathbf{x}_{l,i} \in \mathbb{R}^d$ is the ith input representation at layer l, FF is a two-layer feedforward neural network, $\mathrm{LN}_{\mathrm{out}}$ is a vector-wise layer normalization operation, and $\mathbf{o}_{l,i}$ is the output representation from the multi-head self-attention mechanism, in which H multiple heads simultaneously mix representations from the previous context. This output $\mathbf{o}_{l,i}$ can be decomposed into the sum of representations resulting from each attention head h:

$$\mathbf{o}_{l,i} = \sum_{h=1}^{H} \mathbf{V}_h \begin{bmatrix} \mathbf{X'}_{l,i} \\ \mathbf{1}^\top \end{bmatrix} \mathbf{a}_{l,h,i}, \quad (2)$$

where $\mathbf{X'}_{l,i} \stackrel{\mathrm{def}}{=} [\mathbf{x'}_{l,1}, ..., \mathbf{x'}_{l,i}] \in \mathbb{R}^{d \times i}$ is the sequence of layer-normalized input representations leading up to $\mathbf{x'}_{l,i}$ from the previous context, and $\mathbf{x'}_{l,i} \stackrel{\mathrm{def}}{=} \mathrm{LN}_{\mathrm{in}}(\mathbf{x}_{l,i})$ is the layer-normalized version of $\mathbf{x}_{l,i}$. $\mathbf{V}_h$ represents the head-specific value-output transformation,[1] and $\mathbf{a}_{l,h,i} \in \mathbb{R}^i$ is the vector of attention weights:

$$\mathbf{a}_{l,h,i} = \textsc{SoftMax}\left(\frac{(\mathbf{K}_h \begin{bmatrix} \mathbf{X'}_{l,i} \\ \mathbf{1}^\top \end{bmatrix})^\top \mathbf{Q}_h \begin{bmatrix} \mathbf{x'}_{l,i} \\ 1 \end{bmatrix}}{\sqrt{d_h}} \right), \quad (3)$$

where $\mathbf{Q}_h$ and $\mathbf{K}_h$ represent the head-specific query and key transformations respectively, and $d_h = d/H$ is the dimension of each attention head.

LN_α is a vector-wise layer normalization operation (Ba et al., 2016) that first standardizes the vector and subsequently conducts element-wise transformations using learnable parameters $\mathbf{c}_\alpha, \mathbf{b}_\alpha \in \mathbb{R}^d$:

$$\mathrm{LN}_\alpha(\mathbf{y}) = \frac{\mathbf{y} - m(\mathbf{y})}{s(\mathbf{y})} \odot \mathbf{c}_\alpha + \mathbf{b}_\alpha, \quad (4)$$

where $\alpha \in \{\mathrm{in}, \mathrm{out}\}$, $m(\mathbf{y})$ and $s(\mathbf{y})$ denote the elementwise mean and standard deviation respectively, and $\odot$ denotes a Hadamard product.

[1] As $\begin{bmatrix} \mathbf{W} & \mathbf{b} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} = \mathbf{W}\mathbf{x} + \mathbf{b}$, the bias vectors are omitted from the equations. Additionally, for the simplicity of notation, the 'output' affine transform, which applies to the concatenated 'value-transformed' vectors from all attention heads in a typical implementation, is subsumed into $\mathbf{V}_h$. The bias vector of the 'output' transform is assumed to be distributed equally across heads.

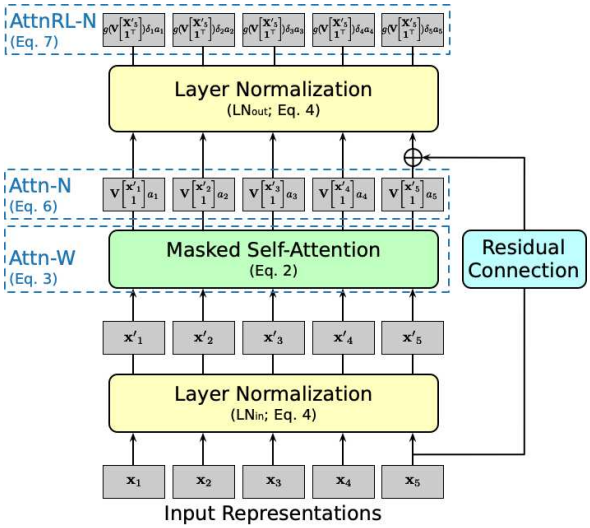

Figure 2: Computations performed within the self-attention block of one head of the GPT-2 language model at a given timestep ($i = 5$). While the masked self-attention mechanism aggregates representations from the previous context in a typical implementation, the linear nature of the subsequent computations allows this aggregation to be deferred to after the residual connection and layer normalization, thereby allowing updated representations to inform alternative formulations of attention weights (i.e. ATTN-N, ATTNRL-N).

2.2 Weight- and Norm-Based Analysis of Previous Context

Previous work that has studied the inner mechanism of Transformer models has focused on analyzing the relative contribution of each token to its final prediction. As a measure that quantifies the 'strength' of contribution, attention weights from the self-attention mechanism have been most commonly used. Similarly to recent work in cognitive modeling (e.g. Ryu and Lewis, 2021), this work also evaluates predictors calculated from attention weights (ATTN-W).

$$\mathbf{a}_{\text{W},l,h,i} = \mathbf{a}_{l,h,i} \tag{5}$$

While analysis using attention weights is common, the assumption that attention weights alone reflect the contribution of each token disregards the magnitudes of the transformed input vectors, as pointed out by Kobayashi et al. (2020). As an alternative, they proposed a norm-based analysis of the self-attention mechanism, which quantifies the contribution of each token as the norm of the transformed vector multiplied by the attention weight. In this work, in order to quantify the relative contribution of each token in the previous context, the norms of the transformed vectors are normalized across the sequence, resulting in 'norm-aware' weights that are comparable to attention weights (ATTN-N).

$$\mathbf{a}_{\text{N},l,h,i} = \frac{\left(\left(\mathbf{V}_h\begin{bmatrix}\mathbf{X}'_{l,i}\\ \mathbf{1}^\top\end{bmatrix} \odot \mathbf{V}_h\begin{bmatrix}\mathbf{X}'_{l,i}\\ \mathbf{1}^\top\end{bmatrix}\right)^\top \mathbf{1}\right)^{\frac{1}{2}} \odot \mathbf{a}_{l,h,i}}{\mathbf{1}^\top\left(\left(\left(\mathbf{V}_h\begin{bmatrix}\mathbf{X}'_{l,i}\\ \mathbf{1}^\top\end{bmatrix} \odot \mathbf{V}_h\begin{bmatrix}\mathbf{X}'_{l,i}\\ \mathbf{1}^\top\end{bmatrix}\right)^\top \mathbf{1}\right)^{\frac{1}{2}} \odot \mathbf{a}_{l,h,i}\right)} \tag{6}$$

More recently, Kobayashi et al. (2021) showed that the residual connection and the layer normalization operation (RESLN; RL) that follow the self-attention mechanism can also be conducted before aggregating representations over token positions. Motivated by this observation, the vector norms that take into consideration these subsequent operations are also examined in this work. Similarly to ATTN-N, the norms are normalized across the sequence to yield weights that are comparable (ATTNRL-N):

$$\mathbf{a}_{\text{RL-N},l,h,i} = \frac{((g(\mathbf{V}_h\begin{bmatrix}\mathbf{X}'_{l,i}\\ \mathbf{1}^\top\end{bmatrix}) \odot g(\mathbf{V}_h\begin{bmatrix}\mathbf{X}'_{l,i}\\ \mathbf{1}^\top\end{bmatrix}))^\top \mathbf{1})^{\frac{1}{2}} \odot \mathbf{a}_{l,h,i}}{\mathbf{1}^\top(((g(\mathbf{V}_h\begin{bmatrix}\mathbf{X}'_{l,i}\\ \mathbf{1}^\top\end{bmatrix}) \odot g(\mathbf{V}_h\begin{bmatrix}\mathbf{X}'_{l,i}\\ \mathbf{1}^\top\end{bmatrix}))^\top \mathbf{1})^{\frac{1}{2}} \odot \mathbf{a}_{l,h,i})}, \tag{7}$$

where $g(\cdot)$ incorporates the residual connection ($+\mathbf{x}_{l,i}$) and layer normalization (LN_{out}) of Eq. 1. Following the assumption that the residual connection serves to 'preserve' the representation at position i (Kobayashi et al., 2021) and that it is distributed equally across heads, $\mathbf{x}_{l,i}$ is added to the representation at position i of each head after dividing it by the number of heads:

$$g(\mathbf{Y}) \overset{\text{def}}{=} \sum_{j=1}^{i}\left(\begin{cases}\frac{\mathbf{Y}\delta_j + \frac{\mathbf{x}_{l,i}}{H\mathbf{a}_{l,h,j[i]}} - m(\mathbf{Y}\delta_j + \frac{\mathbf{x}_{l,i}}{H\mathbf{a}_{l,h,j[i]}})}{s(\mathbf{o}_{l,i}+\mathbf{x}_{l,i})} \odot \mathbf{c}_{\text{out}} + \frac{\mathbf{b}_{\text{out}}}{H} & \text{if } i=j\\ \frac{\mathbf{Y}\delta_j - m(\mathbf{Y}\delta_j)}{s(\mathbf{o}_{l,i}+\mathbf{x}_{l,i})} \odot \mathbf{c}_{\text{out}} + \frac{\mathbf{b}_{\text{out}}}{H} & \text{if } i\neq j\end{cases}\right)\delta_j^\top, \tag{8}$$

where δ_j is a Kronecker delta vector consisting of a one at element j and zeros elsewhere, and $\mathbf{c}_{\text{out}}$ and $\mathbf{b}_{\text{out}}$ refer to the learnable parameters of LN_{out}.

3 Entropy- and Distance-Based Predictors From Attention Patterns

Given the different formulations of self-attention weights, entropy-based predictors that quantify the

diffuseness of self-attention and distance-based predictors that capture the incremental change in attention patterns across timesteps can be defined. The first predictor defined in this work is normalized attention entropy (NAE):

$$\text{NAE}_{\pi,l,h,i} = \frac{\mathbf{a}^{\top}_{\pi,l,h,i[1:i-1]}}{\log_2(i-1)\mathbf{1}^{\top}\mathbf{a}_{\pi,l,h,i[1:i-1]}}(\log_2\frac{\mathbf{a}_{\pi,l,h,i[1:i-1]}}{\mathbf{1}^{\top}\mathbf{a}_{\pi,l,h,i[1:i-1]}}), \quad (9)$$

where $\pi \in \{\text{W}, \text{N}, \text{RL-N}\}$. This is similar to the attention entropy proposed by Ryu and Lewis (2021) as a measure of interference in cue-based recall attributable to uncertainty about the target, with two notable differences. First, NAE controls for the number of tokens in the previous context by normalizing the entropy by the maximum entropy that can be achieved at timestep i. Furthermore, NAE also uses weights over $\mathbf{x'}_{l,1}$, ..., $\mathbf{x'}_{l,i-1}$ that have been re-normalized to sum to 1, thereby adhering closer to the definition of entropy, in which the mass of interest sums to 1.[2]

In addition to NAE, distance-based predictors are defined for capturing effortful change in attention patterns across timesteps. However, as it currently remains theoretically unclear how this distance should be defined, this exploratory work sought to provide empirical results for different distance functions. The first is ΔNAE, which quantifies the change in diffuseness across timesteps:

$$\Delta\text{NAE}_{\pi,l,h,i} = |\text{NAE}_{\pi,l,h,i} - \text{NAE}_{\pi,l,h,i-1}| \quad (10)$$

As with NAE, this predictor is insensitive to how the attention weights are reallocated between tokens in the previous context to the extent that the overall diffuseness remains unchanged.

The second distance-based predictor is Manhattan distance (MD).[3]

$$\text{MD}_{\pi,l,h,i} = ||\mathbf{a}_{\pi,l,h,i} - \mathbf{a}_{\pi,l,h,i-1}||_1 \quad (11)$$

MD directly measures the magnitude of change in attention weights over all tokens at timestep i. MD is less sensitive to the linear distance between tokens and therefore makes it consistent with the predictions of McElree et al. (2003), who found that processing speed was not influenced by the amount of intervening linguistic material in the formation of a dependency.

Finally, the Earth Mover's Distance (EMD; Rubner et al., 2000) is applied to quantify the shift in attention weights. EMD is derived from a solution to the Monge-Kantorovich problem (Rachev, 1985), which aims to minimize the amount of "work" necessary to transform one histogram into another. Formally, let $P = \{(p_1, w_{p_1}), ..., (p_m, w_{p_m})\}$ be the first histogram with m bins, where p_r represents the bin and w_{p_r} represents the weight of the bin; $Q = \{(q_1, w_{q_1}), ..., (q_n, w_{q_n})\}$ the second histogram with n bins; and $\mathbf{D} = [d_{rs}]$ the distance matrix where d_{rs} is the ground distance between bins p_r and q_s. The problem is to find an optimal flow $\mathbf{F} = [f_{rs}]$, where f_{rs} represents the flow between p_r and q_s, that minimizes the overall work.

$$\text{WORK}(P, Q, \mathbf{F}) = \sum_{r=1}^{m}\sum_{s=1}^{n} d_{rs}f_{rs} \quad (12)$$

Once the optimal flow is found, the EMD is defined as the work normalized by the total flow.[4]

$$\text{EMD}(P, Q) = \frac{\sum_{r=1}^{m}\sum_{s=1}^{n} d_{rs}f_{rs}}{\sum_{r=1}^{m}\sum_{s=1}^{n} f_{rs}} \quad (13)$$

To quantify the minimum amount of work necessary to 'transform' the attention weights, the EMD between attention weights at consecutive timesteps is calculated using $\text{EMD}(\mathbf{a}_{\pi,l,h,i-1}, \mathbf{a}_{\pi,l,h,i})$. The ground distance is defined as $d_{rs} = \frac{|r-s|}{i-1}$ in order to control for the number of tokens in the previous context. EMD can be interpreted as being consistent with Dependency Locality Theory (Gibson, 2000) in that reallocating attention weights to tokens further away in the context incurs more cost than reallocating weights to closer tokens.

Code for calculating all of the predictors from GPT-2 under the different attention weight formulations is publicly available at `https://github.com/byungdoh/attn_dist`.

4 Experiment 1: Evaluation of Predictors on Human Reading Times

In order to evaluate the contribution of the entropy- and distance-based predictors, regression models containing commonly used baseline predictors, surprisal predictors, and one predictor of interest were

[2]Preliminary analyses showed that regression models with attention entropy proper without these adjustments failed to converge coherently.

[3]For the purpose of calculating this predictor, the ith element of $\mathbf{a}_{\pi,l,h,i-1}$ is assumed to be 0.

[4]The optimal flow can be found using the transportation simplex method. Additionally, due to the constraint that the total flow is equal to $\min(\sum_{r=1}^{m} w_{p_r}, \sum_{s=1}^{n} w_{q_s})$, the total flow is always equal to 1 in the context of attention weights.

fitted to self-paced reading times and eye-gaze durations collected during naturalistic language processing. In this work, we adopt a statistical procedure that directly models temporal diffusion (i.e. a lingering reponse to stimuli) by estimating continuous impulse response functions and controls for overfitting by assessing the external validity of these predictors through a non-parametric test on held-out data.

4.1 Response Data

The first experiment described in this paper used the Natural Stories Corpus (Futrell et al., 2021), which contains self-paced reading times from 181 subjects that read 10 naturalistic stories consisting of 10,245 words. The data were filtered to exclude observations for sentence-initial and sentence-final words, observations from subjects who answered fewer than four comprehension questions correctly, and observations with durations shorter than 100 ms or longer than 3000 ms. This resulted in a total of 770,102 observations, which were subsequently partitioned into a fit partition of 384,905 observations, an exploratory partition of 192,772 observations, and a held-out partition of 192,425 observations.[5] The partitioning allows model selection (e.g. making decisions about baseline predictors and random effects structure) to be conducted on the exploratory partition and a single hypothesis test to be conducted on the held-out partition, thus obviating the need for multiple trials correction. All observations were log-transformed prior to regression modeling.

Additionally, the set of go-past durations from the Dundee Corpus (Kennedy et al., 2003) also provided the response variable for regression modeling. The Dundee Corpus contains eye-gaze durations from 10 subjects that read 67 newspaper editorials consisting of 51,501 words. The data were filtered to remove unfixated words, words following saccades longer than four words, and words at sentence-, screen-, document-, and line-starts and ends. This resulted in a total of 195,507 observations, which were subsequently partitioned into a fit partition of 98,115 observations, an exploratory partition of 48,598 observations, and a held-out partition of 48,794 observations. All observations were log-transformed before model fitting.

[5]For both datasets, the fit partition, exploratory partition, and held-out partition contain data points whose summed subject and sentence number have modulo four equal to zero or one, two, and three respectively.

4.2 Predictors

For each dataset, a set of baseline predictors that capture basic, low-level cognitive processing were included in all regression models.

- Self-paced reading times (Futrell et al., 2021): word length measured in characters, index of word position within each sentence;
- Eye-gaze durations (Kennedy et al., 2003): word length measured in characters, index of word position within each sentence, saccade length, whether or not the previous word was fixated.

In addition to the baseline predictors, two surprisal predictors were also included in all regression models evaluated in this experiment. The first is unigram surprisal as a measure of word frequency, which was calculated using the KenLM toolkit (Heafield et al., 2013) with parameters estimated on the English Gigaword Corpus (Parker et al., 2009). The second is surprisal from GPT-2 Small (Radford et al., 2019), which is trained on ∼8B tokens of the WebText dataset. Surprisal from the smallest GPT-2 model was chosen because it has been shown to be the most predictive of self-paced reading times and eye-gaze durations among surprisal from all variants of GPT-2 (Oh et al., 2022).

Finally, the entropy- and distance-based predictors defined in Section 3 were calculated from the attention patterns (i.e. $\mathbf{a}_{\pi,l,h,i}$ where $\pi \in \{\text{W}, \text{N}, \text{RL-N}\}$) of heads on the topmost layer of GPT-2 Small. Contrary to previous studies that analyzed the attention patterns of all layers, this work focuses on analyzing those of the topmost layer out of the concern that the attention patterns of lower layers are less interpretable to the extent that they perform intermediate computations for the upper layers. Since the topmost layer generates the representation that is used for model prediction, the attention patterns from this layer are assumed to reflect the contribution of each previous token most directly. Subsequently, the by-head predictors were aggregated across heads to calculate by-word predictors. This assumes that each attention head contributes equally to model prediction, and is also consistent with the formulation of multi-head self-attention in Eq. 2.

To calculate surprisal as well as the entropy- and distance-based predictors, each story of the Natural Stories Corpus and each article of the Dundee Corpus was tokenized according GPT-2's byte-pair encoding (BPE; Sennrich et al., 2016) tokenizer

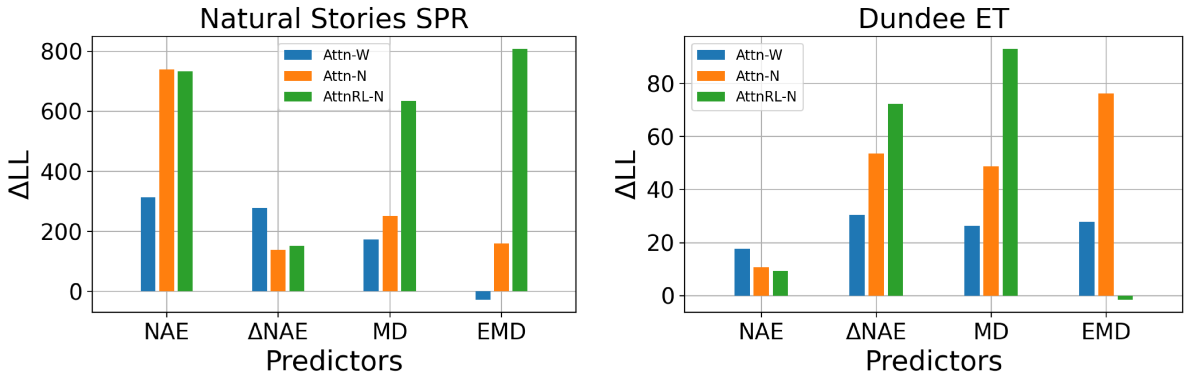

Figure 3: Improvements in CDR model log-likelihood from including each predictor on the exploratory partition of Natural Stories self-paced reading data (left) and Dundee eye-tracking data (right).

and was provided as input to the GPT-2 model. In cases where each story/article did not fit into a single context window, the second half of the previous context window served as the first half of a new context window to calculate predictors for the remaining tokens.[6] Additionally, when a single word w_i was tokenized into multiple subword tokens, negative log probabilities of subword tokens corresponding to w_i were added together to calculate $\mathsf{S}(w_i) = -\log \mathsf{P}(w_i \mid w_{1..i-1})$. Similarly, the entropy- and distance-based predictors for such subword tokens were also added together.

4.3 Procedures

To evaluate the predictive power of each predictor of interest, a *baseline* regression model containing only the baseline predictors and *full* regression models containing one predictor of interest on top of the baseline regression model were first fitted to the fit partition of each dataset. In order to control for the confound of temporal diffusion, continuous impulse response functions (IRFs) were estimated using the statistical framework of continuous-time deconvolutional regression (CDR; Shain and Schuler, 2021). The predictors of interest were centered, and all regression models included a by-subject random intercept.[7]

As a preliminary analysis, the predictive power of different predictors was compared on the exploratory partition by calculating the increase in log-likelihood (ΔLL) to the baseline regression model as a result of including the predictor, following recent work (Goodkind and Bicknell, 2018; Wilcox et al., 2020; Oh et al., 2021). Subsequently, based on the preliminary exploratory results, the predictive power of one best entropy-based predictor and that of one best distance-based predictor were evaluated on the held-out partition of both datasets. More specifically, statistical significance testing was conducted using a paired permutation test (Demšar, 2006) of the difference in by-item squared error between the baseline regression model and the respective full regression models.

4.4 Results

The results in Figure 3 show that across both corpora, most of the entropy- and distance-based predictors make a notable contribution to regression model fit under all attention formulations. Given that the baseline model contains strong predictors such as unigram surprisal and GPT-2 surprisal, this may suggest their validity as predictors of comprehension difficulty. The exception to this is the EMD predictor, which did not show an increase in likelihood under ATTN-W on the Natural Stories Corpus and under ATTNRL-N on the Dundee Corpus. As EMD is sensitive to how the ground distance d_{rs} between bins is defined (i.e. how costly it is to reallocate attention weights across input positions), a more principled definition of d_{rs} may make EMD a more robust predictor.[8] Across the two corpora, ATTN-N+NAE and ATTNRL-N+MD appear to be the most predictive among the entropy- and distance-based predictors respectively.

[6] In practice, most stories/articles fit within two windows.

[7] Early analysis on the exploratory partition revealed that the training data does not support a richer random-effects structure, which led to severe overfitting. Please refer to Appendix A for details on IRF specifications, the optimization procedure of CDR models, as well as the transformations of baseline predictors.

[8] For example, a recency bias may be incorporated by defining the ground distance from tokens closer to the current timestep to be smaller than that from tokens that are farther back in the previous context.

Additionally, the NAE and ΔNAE predictors showed different trends across the two corpora, where NAE contributed to stronger model fit than ΔNAE on the Natural Stories Corpus, while the opposite trend was observed on the Dundee Corpus. In contrast to various surprisal predictors that have shown a very similar trend in terms of predictive power across these two corpora (Oh et al., 2022), these two predictors may shed light on the subtle differences between self-paced reading times and eye-gaze durations. Finally, incorporating vector norms into attention weights (i.e. ATTN-N and ATTNRL-N) generally seems to improve the predictive power of these predictors, which provides support for the informativeness of input vectors in analyzing attention patterns (Kobayashi et al., 2020, 2021).

Table 1 presents the effect sizes of ATTN-N+NAE and ATTNRL-N+MD on the held-out partition of the Natural Stories Corpus and the Dundee Corpus, which were derived by calculating how much increase in reading times the regression model would predict at average predictor value given an increase of one standard deviation. On both datasets, ATTNRL-N+MD appears to be a strong predictor of reading times, which contributed to significantly lower held-out errors. The entropy-based ATTN-N+NAE predictor showed contrasting results across the two corpora, showing a large effect size on the Natural Stories Corpus but not on the Dundee Corpus. This is consistent with the differential results between NAE and ΔNAE on the exploratory partition of the two corpora and may hint at differences between self-paced reading times and eye-gaze durations. In terms of magnitude, the two predictors showed large effect sizes on the Natural Stories Corpus, which were more than twice that of GPT-2 surprisal. On the Dundee Corpus, however, the effect size of ATTNRL-N+MD was much smaller compared to that of GPT-2 surprisal.

Corpus	Predictor	Effect Size (p-value)
Natural Stories	ATTN-N+NAE	6.87 ms ($p < 0.001$)
	GPT2SURP	2.56 ms
	ATTNRL-N+MD	6.59 ms ($p < 0.001$)
	GPT2SURP	2.82 ms
Dundee	ATTN-N+NAE	N/A[9] (n.s.)
	GPT2SURP	4.22 ms
	ATTNRL-N+MD	1.05 ms ($p < 0.001$)
	GPT2SURP	3.81 ms

Table 1: The per standard deviation effect sizes of the predictors on the held-out partition of the Natural Stories Corpus and the Dundee Corpus. Statistical significance was determined by a paired permutation test of the difference in by-item squared error between the baseline regression model and the respective full regression model containing the predictor of interest. The effect sizes of GPT-2 surprisal from the same regression models are presented for comparison.

5 Experiment 2: Do NAE and MD Independently Explain Reading Times?

The previous experiment revealed that on the Natural Stories Corpus, the select entropy- and distance-based predictors from the attention patterns of GPT-2 contributed to significantly higher regression model fit. Although they showed similarly large effect sizes on the held-out partition, the two predictors may independently explain reading times, as they are defined to quantify different aspects of attention patterns. The second experiment examines this possibility following similar procedures as the previous experiment.

5.1 Procedures

In order to determine whether the effect of one predictor subsumes that of the other, a CDR model including *both* ATTN-N+NAE and ATTNRL-N+MD was first fit to self-paced reading times of the fit partition of the Natural Stories Corpus. The CDR model followed the same specifications, random effects structure, and baseline predictors as described in Experiment 1. Subsequently, the fit of this regression model on the held-out partition was compared to those of the two regression models that contain only one of the two predictors from the previous experiment. More specifically, the ΔLL as a result of including the predictor(s) of interest were calculated for each regression model, and statistical significance testing was conducted using a paired permutation test of the difference in by-item squared error between the new 'ATTN-N+NAE & ATTNRL-N+MD' regression model and the respective 'ATTN-N+NAE' and 'ATTNRL-N+MD' regression models, which allowed the contribution of each predictor to be analyzed.

[9]The estimated IRF for ATTN-N+NAE showed a sign error, likely due to poor convergence. Therefore, we treated the effect of this predictor to be statistically non-significant.

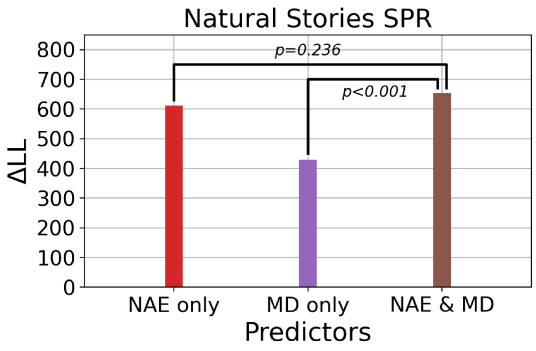

Figure 4: Improvements in CDR model log-likelihood from including ATTN-N+NAE, ATTNRL-N+MD, and both on the held-out partition of Natural Stories self-paced reading data.

5.2 Results

The results in Figure 4 show that the regression model including both ATTN-N+NAE and ATTNRL-N+MD achieves higher ΔLL on the held-out partition of the Natural Stories Corpus compared to regression models including only one of these predictors. Moreover, the difference in by-item squared error between the 'ATTN-N+NAE & ATTNRL-N+MD' regression model and the 'ATTNRL-N+MD' regression model was statistically significant ($p < 0.001$). In contrast, the significance between the 'ATTN-N+NAE & ATTNRL-N+MD' regression model and the 'ATTN-N+NAE' regression model was not statistically significant ($p = 0.236$). This indicates that the entropy-based ATTN-N+NAE contributes to model fit over and above the distance-based ATTNRL-N+MD and also subsumes its effect in predicting reading times.

6 Correlation Analysis Between Predictors and Syntactic Categories

6.1 Procedures

In order to shed more light on the predictors newly proposed in this work, a series of correlation analyses were conducted. First, Pearson correlation coefficients were calculated between the entropy- and distance-based predictors as well as the surprisal predictors to examine the similarity between predictors and the influence of different attention weight formulations. Subsequently, the most predictive ATTN-N+NAE and ATTNRL-N+MD predictors were analyzed with a focus on identifying potential linguistic correlates. This analysis used a version of the Natural Stories Corpus and the Dundee Corpus that had been manually annotated using a generalized categorial grammar annotation scheme (Shain et al., 2018).

6.2 Results

The correlation matrix in Figure 5 shows that within the same predictor, the different attention formulations did not make a very large difference, with 'intra-predictor' correlation coefficients at above 0.85 for most predictors. An exception to this trend was EMD, which showed a correlation coefficient of 0.74 on the Natural Stories Corpus and 0.72 on the Dundee Corpus between ATTN-W+EMD and ATTNRL-N+EMD. Such difference is to be expected, as EMD is the most sensitive to the change of location in attention weights among the distance-based predictors. This is also consistent with the exploratory regression results in Figure 3, where the ΔLL of EMD predictors varied the most as a function of different attention weight formulations.

Additionally, the norm-based attention formulations seem to bring the entropy-based NAE closer to the distance-based predictors in terms of correlation coefficients. On both corpora, ATTN-N+NAE and ATTNRL-N+NAE show stronger correlations to distance-based predictors than ATTN-W+NAE. Interestingly, the highest correlation coefficient between NAE and a distance-based predictor is observed between ATTN-N+NAE and ATTNRL-N+MD at 0.90 on the Natural Stories Corpus and 0.91 on the Dundee Corpus, which were the two strongest predictors identified in Experiment 1. Such high correlation partially explains the results of Experiment 2, in which the influence of ATTN-N+NAE subsumed that of ATTNRL-N+MD.

Finally, the entropy- and distance-based predictors showed moderate correlation to unigram surprisal at around 0.5 on both corpora. With regard to GPT-2 surprisal, these predictors showed weak correlation at around 0.3 on the Natural Stories Corpus, and around 0.4 on the Dundee Corpus. Together with the regression results, this further suggests that the proposed predictors capture a mechanistic process that is distinct from the frequency or predictability of the word.

An analysis of the predictors according to syntactic categories showed that ATTNRL-N+MD may be sensitive to the transition between the subject constituent and the predicate constituent. Its average values for nouns in different contexts presented

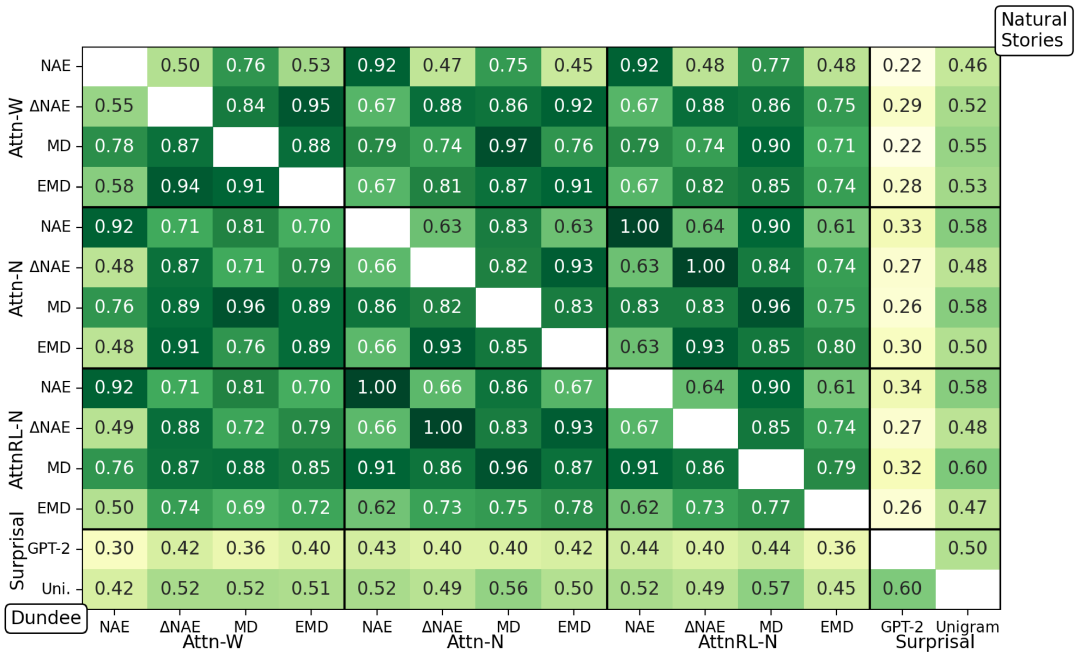

Figure 5: Pearson correlation coefficients between the predictors examined in this work on the Natural Stories Corpus (upper right triangle) and the Dundee Corpus (lower left triangle).

Corpus	Nouns	Count	Mean
Natural Stories	End of subj. NP	1,051	19.024
	Other	3,098	16.520
Dundee	End of subj. NP	4,496	18.982
	Other	15,763	16.661

Table 2: The average ATTNRL-N+MD values for nouns at the end of subject NPs and nouns in other contexts.

in Table 2 show that on both corpora, nouns occurring at the end of subject NPs were associated with a greater shift in attention patterns. This is consistent with the intuition that transitions between constituents entails cognitive effort during incremental processing.

7 Discussion and Conclusion

This work builds on recent efforts to derive memory- and attention-based predictors of processing difficulty to complement expectation-based accounts of sentence processing. Based on the observation that the self-attention patterns of Transformer-based language models can be interpreted as embodying the process of cue-based retrieval, an entropy-based predictor that quantifies the diffuseness of self-attention was first defined. Moreover, based on the idea that reallocation of attention may incur processing difficulty, distance-based predictors that capture the incremental change in attention patterns across timesteps were defined. Regression results using these predictors calculated from the GPT-2 language model showed that these entropy- and distance-based predictors deliver a substantially better fit to self-paced reading and eye-tracking data over a strong baseline including GPT-2 surprisal.

To our knowledge, this work is the first to report robust effects of Transformer attention-based predictors in predicting reading times of broad-coverage naturalistic data. This provides support for Ryu and Lewis's (2021) observation that the self-attention mechanism of Transformers embodies the process of cue-based retrieval, and further suggests that representations that exhibit similarity-based interference can be learned from the self-supervised next-word prediction task. Additionally, the strength of the distance-based predictors further demonstrates the potential to bring together expectation- and memory-based theories of sentence processing under a coherent framework.

Acknowledgments

We thank the reviewers for their helpful comments. This work was supported by the National Science Foundation grant #1816891. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Limitations

The connection between attention patterns of Transformer-based language models and human sentence processing drawn in this work is based on a model trained on English text and data from human subjects that are native speakers of English. Therefore, the connection made in this work may not generalize to other languages. Additionally, although the alternative formulations of self-attention weights resulted in stronger predictors of processing difficulty, they are more computationally expensive to calculate as they rely on an explicit decomposition of the matrix multiplication operation, which are highly optimized in most packages.

Ethics Statement

Experiments presented in this work used datasets from previously published research (Futrell et al., 2021; Kennedy et al., 2003), in which the procedures for data collection and validation are outlined. As this work focuses studying the possible connection between attention patterns of large language models and human sentence processing, its potential negative impacts on society seem to be minimal.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint*.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30.

Timothy Dozat. 2016. Incorporating Nesterov momentum into Adam. In *ICLR Workshop Track*.

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

Edward Gibson. 2000. The Dependency Locality Theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126, Cambridge, MA. MIT Press.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18.

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

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. 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 3543–3556.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7057–7075.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2021. Incorporating residual and normalization layers into analysis of masked language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568.

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

Roger Levy, Evalina Fedorenko, and Edward Gibson. 2013. The syntactic complexity of Russian relative clauses. *Journal of Memory and Language*, 69(4):461–495.

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science*, 10(10):447–454.

Brian McElree, Stephani Foraker, and Lisbeth Dyer. 2003. Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, 48(1):67–91.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22.

Yurii E. Nesterov. 1983. A method for solving the convex programming problem with convergence rate $O\left(\frac{1}{k^2}\right)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547.

Byung-Doh Oh, Christian Clark, and William Schuler. 2021. Surprisal estimators for human reading times need character models. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3746–3757.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5(777963).

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English Gigaword LDC2009T13.

Svetlozar Todorov Rachev. 1985. The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory of Probability and its Applications*, 29(4):647–676.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121.

Soo Hyun Ryu and Richard L. Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71.

Soo Hyun Ryu and Richard L. Lewis. 2022. Using Transformer language model to integrate surprisal, entropy, and working memory retrieval accounts of sentence processing. In *35th Annual Conference on Human Sentence Processing*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Cory Shain and William Schuler. 2021. Continuous-Time Deconvolutional Regression for Psycholinguistic Modeling. *Cognition*, 215.

Cory Shain, Marten van Schijndel, and William Schuler. 2018. Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Workshop on Linguistic and Neuro-Cognitive Resources (LREC 2018)*.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713.

A CDR Implementation Details

The continuous-time deconvolutional regression (CDR) models used in this work were fitted using variational inference to estimate the means and variances of independent normal posterior distributions over all model parameters assuming an improper uniform prior. Convolved predictors used the three-parameter ShiftedGamma impuse response function (IRF) kernel:

$$f(x;\alpha,\beta,\delta) = \frac{\beta^{\alpha}(x-\delta)^{\alpha-1}e^{-\beta(x-\delta)}}{\Gamma(\alpha)} \quad (14)$$

Posterior means for the IRF parameters were initialized at $\alpha = 0.2$, $\beta = 0.5$, and $\delta = -0.2$, which defines a decreasing IRF with a peak centered at $t = 0$ that decays to near-zero within about 1 s. Models were fitted using the Adam optimizer (Kingma and Ba, 2015) with Nesterov momentum (Nesterov, 1983; Dozat, 2016), a constant learning rate of 0.001, and minibatches of size 1,024. For computational efficiency, histories were truncated at 128 timesteps. Prediction from the network used an exponential moving average of parameter iterates with a decay rate of 0.999, and the models were evaluated using *maximum a posteriori* estimates obtained by setting all IRF parameters to their posterior means.

For the baseline regression predictors, the 'index of word position within each sentence' predictors were scaled, and the 'word length in characters' and 'saccade length' predictors were both centered and scaled.