Comparative Evaluation of Computational Models Predicting Eye Fixation Patterns During Reading: Insights from Transformers and Simpler Architectures

Alessandro Lento[1,2], Andrea Nadalini[1], Nadia Khlif[1,3], Vito Pirrelli[1], Claudia Marzi[1] and Marcello Ferro[1,*]

[1]Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale "A. Zampolli", Pisa, Italy
[2]Università Campus Bio-Medico, Roma, Italy
[3]University Mohammed First, Oujda, Morocco

Abstract

Eye tracking records of natural text reading are known to provide significant insights into the cognitive processes underlying word processing and text comprehension, with gaze patterns, such as fixation duration and saccadic movements, being modulated by morphological, lexical, and higher-level structural properties of the text being read. Although some of these effects have been simulated with computational models, it is still not clear how accurately computational modelling can predict complex fixation patterns in connected text reading. State-of-the-art neural architectures have shown promising results, with pre-trained transformer-based classifiers having recently been claimed to outperform other competitors, achieving beyond 95% accuracy. However, transformer-based models have neither been compared with alternative architectures nor adequately evaluated for their sensitivity to the linguistic factors affecting human reading. Here we address these issues by evaluating the performance of a pool of neural networks in classifying eye-fixation English data as a function of both lexical and contextual factors. We show that i) accuracy of transformer-based models has largely been overestimated, ii) other simpler models make comparable or even better predictions, iii) most models are sensitive to some of the major lexical factors accounting for at least 50% of human fixation variance, iv) most models fail to capture some significant context-sensitive interactions, such as those accounting for spillover effects in reading. The work shows the benefits of combining accuracy-based evaluation metrics with non-linear regression modelling of fixed and random effects on both real and simulated eye-tracking data.

Keywords

eye-tracking, eye fixation time prediction, neural network, contextual word embeddings, lexical features

1. Introduction

Eye-tracking records of natural text reading are a valuable window on the cognitive processes underlying word processing and text comprehension. By looking at fixation patterns it is possible to estimate the effects that lexical properties (e.g. length, frequencies, orthographic similarity [1] [2]), contextual constraints (e.g. predictability [3]) and higher-level structures (e.g. syntactic structure or prosodic contour [4]) can have on human word identification and processing. While psycholinguistic experiments have reliably assessed how such effects modulate reading times, it is not clear to what extent computational models of reading can simulate actual behavioural data such as gaze patterns and fixation durations.

Over the past 30 years, research in this field has made considerable progress, leading to the development of sophisticated computational models accounting for fine-grained aspects of eye movement behaviour during word and sentence reading (e.g. EZ-Reader[5], Swift[6]). A significant boost in this area came from large eye-tracking corpora of natural reading (e.g. GECO[7], ZUCO[8], MECO[9]), which allow for (deep) learning models to be tested in prediction tasks of eye tracking metrics. Of late, Hollenstein and colleagues [10] reported that fine-tuned, pre-trained transformer language models can make reliable predictions on a wide range of eye-tracking measurements, covering both early and late stages of lexical processing. The evidence suggests that transformers can inherently encode the relative prominence of language units in a text, in ways that accurately replicate human reading skills and their underlying cognitive mechanisms. Although the accuracy of multilingual transformers is validated across eye-tracking evidence from different languages, the paper neither compares the performance of transformers with the performance of other neural network classifiers trained on the same task, nor it shows what specific knowledge is encoded and put to use by transformers, by looking at the factors affecting their behaviour. In the present paper, we address both issues by assessing the performance of a pool of neural network

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
alessandro.lento@ilc.cnr.it (A. Lento); andrea.nadalini@ilc.cnr.it (A. Nadalini); nadia.khlif@ilc.cnr.it (N. Khlif); vito.pirrelli@ilc.cnr.it (V. Pirrelli); claudia.marzi@ilc.cnr.it (C. Marzi); marcello.ferro@ilc.cnr.it (M. Ferro)
0000-0002-5581-7451 (V. Pirrelli); 0000-0002-3427-2827 (C. Marzi); 0000-0002-1324-3699 (M. Ferro)

classifiers on the English batch of Hollenstein *et al.*'s [10] data.

In what follows, we first describe the English data set and the pool of tested classifiers. Classifiers were selected to include and test either simpler neural architectures than transformers (as is the case with multi-layer perceptrons), or cognitively more plausible processing models (i.e. sequential long-short terms memories). Hybrid models, resulting from the combination of different architectures, were also tested. We then move on to discussing the metrics used in [10] for evaluation, to suggest alternative ways to measure accuracy in a fixation prediction task. Finally, we investigate how sensitive each tested architecture is to a few linguistic factors that are known to account for a sizeable amount of variance in human reading gaze patterns. Although some neural networks turn out to be reasonably good at predicting fixation patterns and replicating some robust psycholinguistic effects that are found in human data, it is still unclear whether this ability is due to specific aspects of their architecture, to the type of information they are provided in input, or to their space of trainable parameters. We conclude that, contrary to recent over-enthusiastic reports, predicting eye-fixation patterns of human natural reading is still a big challenge for currently available neural architectures, including transformer-based ones. For this very reason, we contend that the task is key to understanding the inductive bias of these models, as well as assessing their cognitive plausibility as models of language behaviour.

2. Data and Experiments

All models described in the following paragraphs were trained, validated, and tested on data from the GECO corpus [7]. We used a 5-fold cross-validation with 95% training, 5% validation and 5% test. Experiments were conducted using the PyTorch library [11] in Python or MatLab [12].

2.1. Dataset

The GECO corpus [7] contains data from 14 English native speakers whose eye movements were recorded while reading Agatha Christie's novel "The Mysterious Affair at Styles" (56410 tokens). Out of the eight word-level eye tracking measurements used in [10], we focused on i) first-pass duration (FPD) (the time spent fixating a word the first time it is encountered, averaged over subjects, see Fig. 2) and ii) fixation proportion (FPROP) or probability (number of subjects that fixated a word, divided by the total number of subjects).

Word tokens in the original dataset were encoded with linguistic information including:

i) character length (removing punctuation)

ii) log frequency (source: BNC [13])

iii) part-of-Speech tag (source: Stanza [14])

iv) context surprisal/predictability (source: GPT-2 [15, 16, 3])

v) distance from the beginning of the sentence (number of intervening tokens)

vi) distance from the end of the sentence (number of intervening tokens)

vii) presence of heavy punctuation after the token

viii) presence of light punctuation after the token.

2.2. BERT ++

To replicate results from [10], we used BERT [17] with a linear layer on top of it. The linear layer gets BERT **contextual word embeddings** as input, to predict FPD and FPROP.

After sentence padding and tokenization, irrelevant and special subtokens were masked to enforce a correspondence between each vector in the target sequence and each vector in the output sequence, and train the loss only on relevant tokens. Mean Square Error (MSE) loss was used along with the AdamW optimizer (with no weight decay for the biases). The initial learning rate was set to $5 \cdot 10^{-5}$, and a linear scheduler was used. We used a 16 sentences batch size and 100 training epochs, with an early stopping criterion (best model on the validation set). The model was trained both with fine-tuning (i.e. by also training BERT internal weights: **bert FT + layer**) and without fine-tuning (by only training final layer weights: **bert + layer**).

Finally, we used BERT also in combination with a sequential LSTM network. This model (**bert + LSTM**) takes the pre-trained BERT **contextual word embeddings** (i.e. without fine-tuning) in input, along with the lexical features (i), (ii) and (iv), to predict FPD and FPROP.

2.3. LSTM

Reading is inherently sequential. Thus, recurrent neural networks appear to offer a promising approach to modelling a fixation prediction task, and a good alternative to transformers. Using the GECO dataset split into pages rather than sentences, we trained an LSTM with 96 hidden units and a single layer, with a feed-forward network using *tanh* activation functions on top of it. The model (**lstm**) takes as input the lexical features (i)-(iv) for the target token and 4 tokens to its left and 3 to its right, to predict FPD and FPROP of the target token. MSE loss was used along with the AdamW optimizer. The initial learning rate was set to $5 \cdot 10^{-3}$, with a linear scheduler

and a batch containing the entire training dataset. The model was trained for 3000 epochs with an early stopping criterion (best model on the validation set).

2.4. MLP

A Multi-Layer-Perceptron (**mlp**) was trained using the entire set of lexical features (i)-(viii) as input, with an input context consisting of the two words immediately preceding and ensuing the target word. Several instances of this architecture were tested, but only the results of the best performing instance (with a single hidden layer of 10 units, sigmoidal activation functions, the Adam optimiser, the MSE loss, a constant learning rate of 0.1, and 1000 training epochs) are reported here.

An identical MLP model (**mlp UDT**) was eventually trained on a subset of GECO training data, obtained by sampling target features uniformly. This was done to train the network with an equal number of tokens for each bin of fixation times, and assess the impact of different distributions of input data on the network's performance on test data.

2.5. Evaluation

We evaluated the performance of all our models using three accuracy metrics based on the absolute error between the predicted value o_i and the target value t_i on the *i-th* token of the GECO dataset:

$$e_i = |o_i - t_i|$$

Loss accuracy (**accL**) is a measure of the overall similarity between predicted and target values, calculated as the complement to 1 of the Mean Absolute Error (MAE) after fitting the target data t_i in the training set into the [0; 1] range with the *min-max* scaling:

$$accL(set) = 1 - \frac{1}{N_{set}} \sum_{i \in set}^{N_{set}} \hat{e}_i$$

where $\hat{e}_i = |\hat{o}_i - \hat{t}_i|$, $\hat{t}_i = t_i / \max_{j=trainingset}\{t_j\}$, and $\hat{o}_i$ is the model prediction for $\hat{t}_i$. *Loss accuracy* is the metric used in [10].

Threshold accuracy (**accT**) measures how many times the predicted value is close to the target value within a fixed threshold, and is calculated as follows:

$$accT(set) = 1 - \frac{1}{N_{set}} \sum_{i \in set}^{N_{set}} \theta[e_i - \epsilon]$$

Sensitivity accuracy (**accS**) counts how many times the predicted value is close to the target value within a threshold dynamically calculated on the basis of the target value: the higher the target value, the higher the threshold. An offset value is needed to obtain a positive threshold also for zero target values. This is calculated as follows:

$$accS(set) = 1 - \frac{1}{N_{set}} \sum_{i \in set}^{N_{set}} \theta\left[e_i - (\alpha \cdot t_i + \epsilon)\right]$$

where N_{set} is the number of examples in the training/test set, θ is the Heaviside step function, ϵ is a threshold and α is a sensitivity coefficient.

As for FPD, which is a duration expressed in seconds, we used $\epsilon = 25ms$ and $\alpha = 10\%$ for *accS*, and $\epsilon = 50ms$ for *accT*. As for FPROP, which is a probability, we used $\epsilon = 0.01$ and $\alpha = 10\%$ for *accS*, and $\epsilon = 0.1$ for *accT*.

Finally, the performance of our models was compared against a baseline model (**const**) that always outputs the overall mean fixation duration (across both subjects and items) in the training data.

3. Results

Models' results for FPD prediction are summarised in Table 1 and plotted in Fig. 1. The *accL* results reported in [10] for **bert FT + layer** are essentially replicated. However, being a simple average over all test instances, **accL** is blind to error magnitude, as well as the possible presence of prediction biases for specific ranges of fixation values. Note that the **const** model, which predicts the same average FDP for every token in the test set, scores a flattering 95.68% on **accL**, vs. 36.97% on *accS*, and 48.10% on *accT*

Table 2 summarises *accS* values of all models, by binning them into three FPD ranges.

4. Data analysis

To what extent are neural network models sensitive to some of the factors accounting for gaze patterns in human natural reading? Are language models able to adapt themselves to both lexical properties and in-context features of a reading text, thus exhibiting a human-like performance?

Human reading behaviour is shown to be affected by lexical features – e.g. word length and frequency, and morphological complexity – as well as by contextual factors, with a facilitatory effect of contextual redundancy and predictability (18, 19) on reading duration and eye fixations. Accordingly, we modelled human FPDs as a response variable resulting from the interaction of both lexical and contextual predictors: namely, word length, a dichotomous classification of token POS into content versus function words, surprisal of the target word as a

	FPD accuracies					
	test			training		
model	*accS*	*accT*	*accL*	*accS*	*accT*	*accL*
const	36.97%	48.10%	95.68%	37.07%	48.06%	95.69%
	(0.83%)	(1.00%)	(0.05%)	(0.04%)	(0.05%)	(0.00%)
bert	*55.02%*	67.82%	97.05%	*58.11%*	*70.74%*	97.25%
+ layer	(0.86%)	(0.99%)	(0.05%)	(0.82%)	(0.70%)	(0.05%)
mlp UDT	56.41%	*67.79%*	96.21%	61.21%	72.37%	96.52%
	(0.35%)	(0.79%)	(1.25%)	(0.95%)	(0.57%)	(1.08%)
bert	58.49%	70.01%	*95.38%*	63.64%	75.89%	*95.90%*
+ lstm	(0.91%)	(0.82%)	(0.07%)	(0.48%)	(0.77%)	(0.97%)
bert FT	57.80%	70.03%	97.23%	**93.18%**	**94.81%**	**98.80%**
+ layer	(1.02%)	(1.13%)	(0.05%)	(0.81%)	(0.71%)	(0.05%)
mlp	**60.16%**	73.05%	**97.39%**	60.63%	73.31%	97.40%
	(0.85%)	(0.78%)	(0.04%)	(0.37%)	(0.24%)	(0.01%)
lstm	60.01%	**73.18%**	**97.39%**	61.66%	74.27%	97.45%
	(0.38%)	(0.31%)	(0.03%)	(0.24%)	(0.19%)	(0.01%)

Table 1
Overall FPD prediction accuracy in the GECO dataset. For each model, three different accuracy scores are given as described in the text; **const** is used as a baseline; highest accuracies in bold; lowest accuracies in italics.

	3-bin FPD accuracy on test		
model	*low*	*medium*	*high*
const	0.00%	41.08%	0.00%
bert + layer	21.43%	58.98%	*23.02%*
mlp UDT	**52.33%**	*56.91%*	**51.49%**
bert + lstm	24.19%	62.17%	26.61%
bert FT + layer	32.86%	62.65%	31.65%
mlp	*11.77%*	**64.38%**	32.62%
lstm	19.05%	64.26%	29.45%

Table 2
Sensitivity accuracy (*accS*) values for three bins from the FPD distribution: *low* (FPD below the 5^{th} percentile = 36ms), *medium* (FPD ranging from the 5^{th} to the 95^{th} percentile), and *high* (FPD above the 95^{th} percentile = 280ms).

measure of how unexpected or unpredictable the word is, and the probability of the word immediately preceding the target word in context (to account for so-called spill-over effects). Additionally, we used a *Generalised Additive Model (GAM)*, with token log-frequency as a smooth term, to model for possibly non-linear effects of predictors. Models' coefficients and effect plots are shown in Appendix C (Figure 3 and Table 4).

GAMs with identical independent variables have been run to model the FPDs predicted by all our neural networks, on both training and test data. Inspection of effect plots and model coefficients – as reported in Appendix C – shows a behavioural alignment of all models with human data for what concerns the modulation of fixation times by lexical features, in both train and test data.

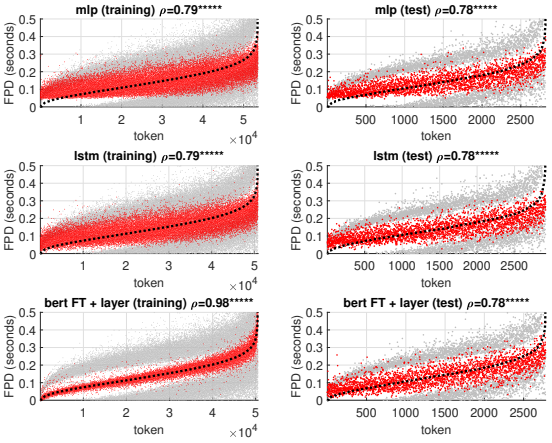

Figure 1: Models predictions (red dots) plotted with target FPD values (black dots), after ordering tokens for increasing FPDs. Grey dots represent averaged FPD values plus\minus their standard deviation across participants. Left: training data. Right: test data. From top to bottom: MLP, LSTM, BERT fine-tuned. For each plot, the Spearman-ρ correlation coefficient between predicted and target values is shown along with the significance value.

In contrast, all models fail to capture some contextual effects on test data, such as those observed in a context window of – at least – two adjacent words. To illustrate, efficient syntactic chunking (e.g. of noun, verb and prepositional phrases) has been shown to lead to faster and more accurate human reading (see, for example, [20]). Conversely, most neural networks show no statistically significant effect on fixation duration of the probability of the immediately preceding word in context. This is observed either is isolation (probMinus1) in LSTMs and transformer-based models with BERT representations (either fine-tuned or not), or in interaction with the unpredictability of the target word (surprisal:probMinus1). The evidence shows that most neural models cannot replicate, among other things, so-called *spillover* effects of the left-context on the reading time of ensuing words [21].

5. General Discussion

Transformer-based neural networks appear to reasonably predict fixation probability and first-pass duration of words in human reading of English connected texts. Our present investigation basically supports this conclusion, while providing new evidence on two related questions. Two questions naturally arise in this context. How accurate are transformer-based predictions compared with the best predictions of other neural network classifiers trained on the same task? How cognitively plausible are the mechanisms underpinning this perfor-

mance? Here, we addressed both questions by testing various models on the task of predicting human reading measurements from the GECO corpus, using different evaluation metrics and regressing network predictions on a few linguistic factors that are known to account for human reading behaviour.

Our first observation is that assessing a network's performance by looking at its MAE loss function provides a rather gross evaluation of the effective power of a neural network simulating human reading behaviour. A baseline model assigning each token a constant gaze duration that equals the average of all FPD values attested in GECO achieves a 95.7% loss-based accuracy on both test and training data. That a transformer-based classification scores 97.2% on the same metric and the same test data cannot be held, as such, as a sign of outstanding performance. In fact, it turns out that the MAE loss function is blind to both the magnitude of a network error, and possible biases in the prediction of very low/high target values. Thus, it provides an inflated estimate of a model's accuracy. We suggest that binary evaluation metrics, based on a fixed threshold partially overcome these limitations. Yet, as single word fixation times typically range between tens to hundreds of milliseconds, application of a fixed threshold will differently affect tokens with different fixation times. We conclude that a relative threshold based on each word's fixation time is a fairer way to measure prediction accuracy. Clearly, this comes at a cost. When assessed with a relative threshold, the accuracy of a transformer-based architecture on test data drops from 70% down to 57.8%.

It turned out that all other network models tested for the present purposes showed accuracy levels that are comparable to the accuracy of a transformer-based architecture. Since the former are trained on a more restricted set of lexical and contextual input features than the latter, this seems to suggest that word embeddings are of limited use in the task at hand. Although fine-tuned word embeddings actually appear to score much higher on training data (even using *accT* and *accS*), we observe that this is due to data overfitting, as clearly shown by the considerably poorer performance of the fine-tuned model on test data.

An analysis of the psychometric plausibility of the gaze patterns simulated with our neural models reveals that a relatively small set of linguistic factors that are known to account for a sizeable amount of variance in human fixation times can also account for the bulk of variance in models' behaviour. This is relatively unsurprising, as most of these models were trained on input features that encode at least some of these factors. Nonetheless, we believe that the result is interesting for at least two reasons. First, it shows a promising convergence between computational metrics of model accuracy and quantitative models of psychometric assessment. Secondly, it suggests that one can gain non trivial insights in a model's behaviour by analysing to what extent the behaviour is sensitive to the same linguistic factors human readers are known to be sensitive to. On the one hand, this is a step towards understanding what information a neural model is actually learning and putting to use for the task. On the other hand, this is instrumental in developing better models, as it shows what type of input information is more needed to successfully carry out a task, at least if one is trying to simulate the way the same task is carried out by speakers.

In the end, it may well be the case that a 70% fixed-threshold accuracy in simulating average gaze patterns in human reading is not as disappointing as it might seem. Given the wide variability in human reading behaviour (and even in a single reader when confronted with different texts), a considerable amount of variance in our data may simply be accounted for by by-subject (or by-token) random effects. In some experiments not reported here we trained our models to predict single-reader behaviour. All architectures fared rather poorly on the task, a result which is in line with similar disappointing results on other output features reported in [10]. Looking back at Figure 1, it can be noted that all models' predictions fall into a $\mu_i \pm \sigma_i$ range, where μ_i and σ_i are, respectively, the by-reader mean and standard deviation of FPD values for token i (see also Table 2). This pattern may suggest that models' predictions are in fact bounded by the standard deviation we observe in human behaviour and cannot reach out of these bounds. Conversely, this evidence may be interpreted as suggesting that more input features are needed to build more accurate classifiers. Further experiments are needed to test the merits of either conjecture.

6. Limitations and outlook

In the present paper, we replicated recent experimental data of transformer-based architectures simulating word fixation duration in reading a connected text [10], with a view to assessing their relative performance compared with reading times by humans and other neural architectures. This justifies our exclusive focus on fixation duration, which is, admittedly, only one behavioural correlate of a complex, inherently multimodal task such as reading. In fact, reading requires the fine coordination of eye movements and articulatory movements for text decoding and comprehension. The eye provides access to the visual stimuli needed for voice articulation to unfold at a relatively constant rate. In turn, articulation can feedback oculomotor control for eye movements to be directed when and where processing difficulties arise. Incidentally, this is also true of silent reading as shown by evidence supporting the Implicit Prosody Hypothesis

[22], i.e. the idea that, in silent reading, readers activate prosodic representations that are similar to those they would produce when reading the text aloud. Hence, a reader must always rely on a tight control strategy to ensure that fixation and articulation are optimally coordinated.

A clear limitation of our current work and all experiments reported here is that we are only focusing on one dimension of a complex, multimodal behaviour like reading. Recently, we showed that there is a lot about gaze patterns that we can understand by correlating eye movements with voice articulation [23]. This information, which cannot be represented in a dataset structured at the word level, may be critical for a model to accurately learn and mimic the cognitive mechanisms underlying natural reading. Likewise, as correctly pointed out by one of our reviewers, focusing on fixation times while ignoring saccadic movements may seriously detract from the explanatory power of any computational model of human reading. In fact, this could be tantamount to timing a bike rider's speed, while ignoring if she is climbing up a hill or approaching a sharp turn. More realistic models of reading are bound to include more aspects of reading behaviour in more ecologically valid tasks. In the end, it may well be the case that the task of predicting gaze patterns of human reading should be conceptualized differently, by anchoring these patterns not only to the syntagmatic dimension of a written text, but also to the time-line of the different movements and multimodal processes that unfold during reading.

Acknowledgments

The present study has partly been funded by the *ReadGround* research grant from the National Research Council (CNR), and the *ReMind* and *Braillet* PRIN grants, from the Ministry of University and Research (MUR).

Alessandro Lento is a PhD student enrolled in the *National PhD in Artificial Intelligence*, XXXVII cycle, course on Health and Life sciences, organized by Università Campus Bio-Medico in Rome.

Nadia Khlif is a PhD student in the *Computer Science Research Laboratory*, Faculty of Sciences, at the University Mohammed First of Oujda, Morocco.

Andrea Nadalini's work is kindly covered by the "RAISE - Robotics and AI for Socio-economic Empowerment" grant (ECS00000035), funded by the European Union - NextGenerationEU and by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.5.

References

[1] S. Gerth, J. Festman, Reading development, word length and frequency effects: An eye-tracking study with slow and fast readers, Frontiers in Communication 6 (2021) 743113.

[2] S. Schroeder, T. Häikiö, A. Pagán, J. H. Dickins, J. Hyönä, S. P. Liversedge, Eye movements of children and adults reading in three different orthographies., Journal of Experimental Psychology: Learning, Memory, and Cognition 48 (2022) 1518.

[3] L. Salicchi, E. Chersoni, A. Lenci, A study on surprisal and semantic relatedness for eye-tracking data prediction, Frontiers in Psychology 14 (2023) 1112365.

[4] M. Hirotani, L. Frazier, K. Rayner, Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements, Journal of Memory and Language 54 (2006) 425–443.

[5] E. D. Reichle, K. Rayner, A. Pollatsek, The E-Z Reader model of eye-movement control in reading: Comparisons to other models, Behavioral and Brain Sciences 26 (2003) 445–476.

[6] R. Engbert, A. Nuthmann, E. Richter, R. Kliegl, SWIFT: A Dynamical Model of Saccade Generation During Reading., Psychological review 112 (2005) 777–813.

[7] U. Cop, N. Dirix, D. Drieghe, W. Duyck, Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading, Behavior Research Methods 49 (2017) 602–615.

[8] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, N. Langer, ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading, Scientific Data 5 (2018) 180291.

[9] N. Siegelman, S. Schroeder, C. Acartürk, H.-D. Ahn, S. Alexeeva, S. Amenta, R. Bertram, R. Bonandrini, M. Brysbaert, D. Chernova, S. M. Da Fonseca, N. Dirix, W. Duyck, A. Fella, R. Frost, C. A. Gattei, A. Kalaitzi, N. Kwon, K. Lõo, M. Marelli, T. C. Papadopoulos, A. Protopapas, S. Savo, D. E. Shalom, N. Slioussar, R. Stein, L. Sui, A. Taboh, V. Tønnesen, K. A. Usal, V. Kuperman, Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO), Behavior Research Methods 54 (2022) 2843–2863.

[10] N. Hollenstein, F. Pirovano, C. Zhang, L. Jäger, L. Beinborn, Multilingual language models predict human reading behavior, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 106–123.

[11] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Consta-

ble, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. K. Luk, B. Maher, Y. Pan, C. Puhrsch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, S. Zhang, M. Suo, P. Tillet, X. Zhao, E. Wang, K. Zhou, R. Zou, X. Wang, A. Mathews, W. Wen, G. Chanan, P. Wu, S. Chintala, PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation, in: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, volume 2 of *ASPLOS '24*, Association for Computing Machinery, 2024, pp. 929–947.

[12] T. M. Inc., Matlab version: 9.7.0.1190202 (r2019b), 2019.

[13] B. Consortium, The british national corpus, xml edition, 2007.

[14] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).

[16] J. A. Michaelov, B. K. Bergen, Do language models make human-like predictions about the coreferents of italian anaphoric zero pronouns?, arXiv preprint arXiv:2208.14554 (2022).

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. ArXiv:1810.04805 [cs] version: 2.

[18] K. E. Stanovich, Attentional and automatic context effects in reading, in: Interactive processes in reading, Routledge, 2017, pp. 241–267.

[19] G. B. Simpson, R. R. Peterson, M. A. Casteel, C. Burgess, Lexical and sentence context effects in word recognition., Journal of Experimental Psychology: Learning, Memory, and Cognition 15 (1989) 88.

[20] K. Rayner, K. H. Chace, T. J. Slattery, J. Ashby, Eye movements as reflections of comprehension processes in reading, Scientific studies of reading 10 (2006) 241–255.

[21] N. J. Smith, R. Levy, The effect of word predictability on reading time is logarithmic, Cognition 128 (2013) 302–319.

[22] M. Breen, Empirical investigations of the role of implicit prosody in sentence processing, Language and Linguistics Compass 8 (2014) 37–50.

[23] A. Nadalini, C. Marzi, M. Ferro, L. Taxitari, A. Lento, D. Crepaldi, V. Pirrelli, Eye-voice and finger-voice spans in adults' oral reading of connected texts. Implications for reading research and assessment, The Mental Lexicon (2024). URL: https://benjamins.com/catalog/ml.00025.nad.

[24] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2023. URL: https://www.R-project.org/.

A. GeCO FPD data

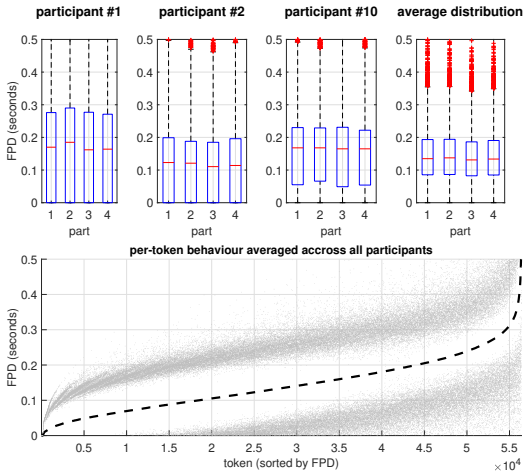

Figure 2: A view of FPD data in the GECO dataset, consisting of eye-tracking patterns of 14 adult participants reading the novel "The Mysterious Affair at Styles" by Agata Christie. **Top panel:** distributions of FPD data, with chapters grouped into 4 parts, for participant #1 (with 3 more participants showing a similar distribution), participant #2 (with 8 more participants showing a similar distribution) and participant #10. The rightmost box plot shows the average distribution across all 14 participants. **Bottom panel:** plot of all 56410 tokens in the dataset, in ascending order of mean FPD (dashed black line). For each token, the standard deviation calculated on the distribution of the FPDs of the 14 participants is shown both above and below the mean value (gray dots).

B. FPROP accuracy

	FPROP accuracies					
	test			training		
model	*accS*	*accT*	*accL*	*accS*	*accT*	*accL*
const	2.70%	7.17%	51.44%	2.82%	7.37%	51.71%
	(0.37%)	(0.70%)	(0.57%)	(0.02%)	(0.04%)	(0.03%)
bert	*33.84%*	*44.86%*	*86.34%*	*37.47%*	*48.84%*	*87.68%*
+ layer	(1.28%)	(0.89%)	(0.15%)	(1.24%)	(1.24%)	(0.28%)
mlp UDT	36.24%	48.75%	86.90%	43.40%	58.64%	89.49%
	(0.37%)	(0.83%)	(0.21%)	(0.71%)	(0.61%)	(0.09%)
bert	38.00%	48.46%	87.50%	42.78%	54.78%	89.16%
+ lstm	(0.76%)	(1.01%)	(0.43%)	(0.88%)	(0.70%)	(0.12%)
bert FT	36.39%	47.60%	87.00%	**75.10%**	**90.66%**	**95.28%**
+ layer	(1.09%)	(1.23%)	(0.33%)	(1.78%)	(1.85%)	(0.26%)
mlp	**38.96%**	**51.23%**	**88.10%**	39.45%	51.78%	88.34%
	(1.05%)	(1.08%)	(0.19%)	(0.27%)	(0.15%)	(0.02%)
lstm	37.91%	49.95%	87.93%	39.42%	51.63%	88.34%
	(0.85%)	(0.78%)	(0.11%)	(0.46%)	(0.42%)	(0.12%)

Table 3
Accuracy values of neural models predicting the fixation probabilities of the GECO dataset. For each model three different accuracy metrics are used, as described in the paper. The "const" model was used as a baseline; highest accuracy scores are highlighted in bold: lowest scores are shown in italic

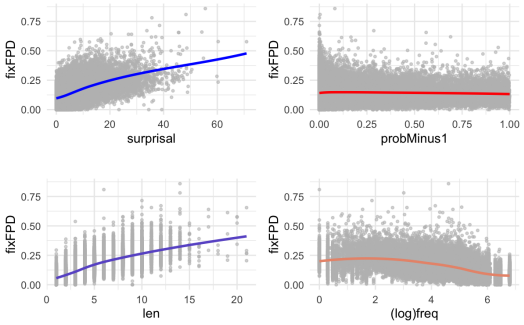

Figure 3: Effects of surprisal, probability of the preceding token (*probMinus1*), word length (*len*) as predictors, and word log-frequency (*logFreq*) as a smooth term, on human fixation first-pass duration (fixFPD) as a response variable.

C. Data analysis

In this section, coefficients of Generalised Additive Models (GAMs) are detailed for each neural model. Statistical non-significant *p-values* on GAM predicting terms are given in bold-face. GAMs are fitted using the package `gamm4` version 0.2-6 of the *R* statistical software [24], as they do not assume a linear relation between the fitted variable and its predictors. All plots were created via the `ggplot2` package, version 3.5.

	Human FPD			
parametric coeff.	estimate	std. error	t value	pr(>|t|)
Intercept (content)	6.960e-02	7.858e-04	88.568	$< 2e-16$
surprisal	1.928e-03	5.002e-05	38.539	$< 2e-16$
probMinus1	-1.395e-02	1.363e-03	-10.233	$< 2e-16$
Intercept (function)	-2.599e-02	1.143e-03	-22.746	$< 2e-16$
length (content)	1.562e-02	1.423e-04	109.767	$< 2e-16$
length (function)	5.499e-03	2.791e-04	19.704	$< 2e-16$
surprisal:probMinus1	4.692e-04	1.776e-04	2.642	< 0.01
s(logFreq)				$< 2e-16$
R^2	58.4%			

Table 4
GAM coefficients fitting human fixation FPD: FPD $\sim$ surprisal $\times$ probMinus1 $+$ POSgroup $\times$ wordlength $+$ s(logFreq).

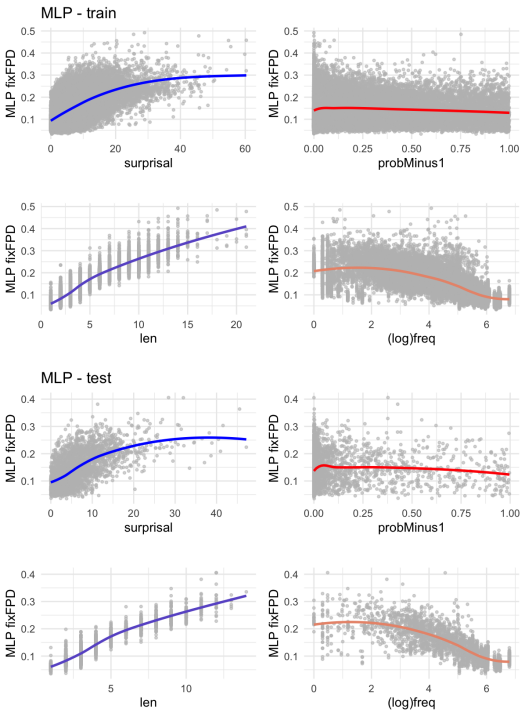

Figure 4: MLP effects in training (**top panel**) and test (**bottom panel**) data, with surprisal, probability of the preceding token (*probMinus1*), word length (*len*) as predictors, word log-frequency as a smooth term (*logFreq*), and fixation first-pass duration as response variable.

	MLP FPD			
parametric coeff.	estimate	std. error	t value	pr(>|t|)
Intercept (content)	7.252e-02	2.729e-04	265.71	$< 2e-16$
surprisal	9.028e-04	1.734e-05	52.064	$< 2e-16$
probMinus1	-1.417e-02	4.723e-04	-29.995	$< 2e-16$
Intercept (function)	-2.312e-02	3.973e-04	-58.2006	$< 2e-16$
length (content)	1.651e-02	4.935e-05	334.512	$< 2e-16$
length (function)	4.324e-03	9.698e-05	44.584	$< 2e-16$
surprisal:probMinus1	1.810e-04	6.166e-05	2.936	< 0.005
s(logFreq)				$< 2e-16$
R^2	92.2%			
Intercept (content)	7.148e-02	1.183e-03	60.42	$< 2e-16$
surprisal	7.585e-04	7.619e-05	9.956	$< 2e-16$
probMinus1	-1.061e-02	2.044e-03	-5.188	$< 2.2e-07$
Intercept (function)	-1.919e-02	1.658e-03	-11.573	$< 2e-16$
length (content)	1.677e-02	2.136e-04	78.502	$< 2e-16$
length (function)	3.399e-03	3.963e-04	8.5774	$< 2e-16$
surprisal:probMinus1	-1.408e-04	2.480e-04	-0.568	**0.57**
s(logFreq)				$< 2e-16$
R^2	92.6%			

Table 5
GAM coefficients fitting MLP fixation FPD in training (**top**) and test (**bottom**) data: FPD $\sim$ surprisal $\times$ probMinus1 $+$ POSgroup $\times$ wordlength $+$ s(logFreq).

	LSTM FPD			
parametric coeff.	estimate	std. error	t value	pr(>|t|)
Intercept (content)	7.051e-02	3.259e-04	216.317	$< 2e-16$
surprisal	7.615e-04	2.069e-05	36.802	$< 2e-16$
probMinus1	2.120e-03	5.644e-04	3.756	< 0.001
Intercept (function)	-1.600e-02	4.778e-04	-33.492	$< 2e-16$
length (content)	1.649e-02	5.896e-05	279.739	$< 2e-16$
length (function)	2.801e-03	1.170e-04	23.945	$< 2e-16$
surprisal:probMinus1	-3.385e-04	7.325e-05	-4.621	< 0.001
s(logFreq)				$< 2e-16$
R^2	89.6%			
Intercept (content)	6.812e-02	1.407e-03	48.431	$< 2e-16$
surprisal	6.837e-04	9.284e-05	7.364	$< 2.3e-13$
probMinus1	3.293e-03	2.458e-03	1.340	**0.18**
Intercept (function)	-1.255e-02	1.936e-03	-6.480	$< 1.1e-10$
length (content)	0.0152041	0.0004032	37.709	$< 2e-16$
length (function)	0.0042481	0.0007472	5.685 1	$< 1.4e-08$
surprisal:probMinus1	-0.0001970	0.0004701	-0.419	**0.67**
s(logFreq)				$< 2e-16$
R^2	89.9%			

Table 6
GAM coefficients fitting LSTM fixation FPD in training (**top**) and test (**bottom**) data: FPD $\sim$ surprisal $\times$ probMinus1 $+$ POSgroup $\times$ wordlength $+$ s(logFreq).

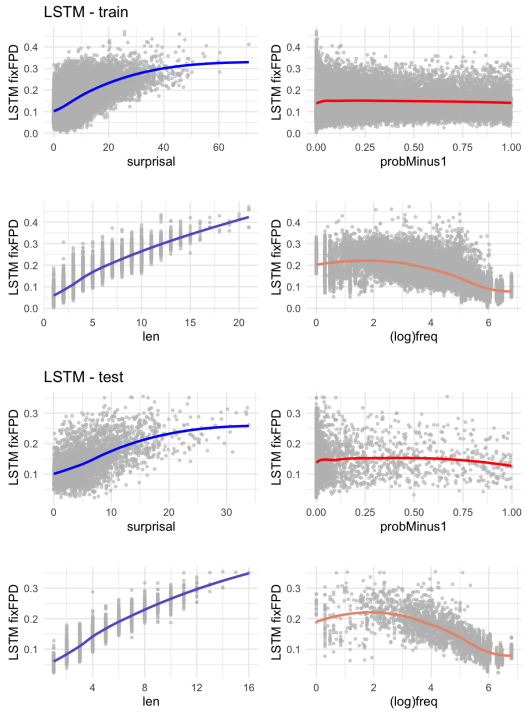

Figure 5: LSTM effects in training (**top panel**) and test (**bottom panel**) data, with surprisal, probability of the preceding token (*probMinus1*), word length (*len*) as predictors, word log-frequency as a smooth term (*logFreq*), and fixation first-pass duration as response variable.

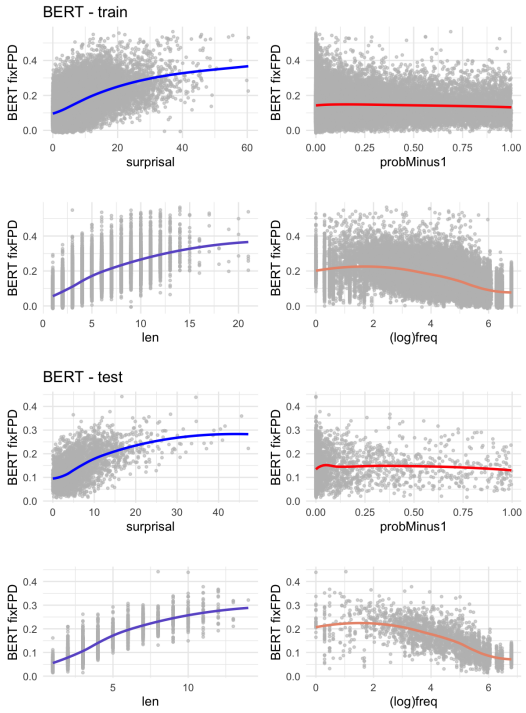

Figure 6: fine-tuned BERT effects in training (**top panel**) and test (**bottom panel**) data, with surprisal, probability of the preceding token (*probMinus1*), word length (*len*) as predictors, word log-frequency as a smooth term (*logFreq*), and fixation first-pass duration as response variable.

	BERT+fine-tuning FPD			
parametric coeff.	estimate	std. error	t value	pr(>|t|)
Intercept (content)	6.950e-02	8.572e-04	81.075	$< 2e-16$
surprisal	2.013e-03	5.446e-05	36.9562	$< 2e-16$
probMinus1	-1.475e-02	1.483e-03	-9.9416	$< 2e-16$
Intercept (function)	-2.631e-02	1.248e-03	-21.0852	$< 2e-16$
length (content)	1.570e-02	1.550e-04	101.307	$< 2e-16$
length (function)	5.528e-03	3.046e-04	18.148	$< 2e-16$
surprisal:probMinus1	5.024e-04	1.937e-04	2.594	< 0.01
s(logFreq)				$< 2e-16$
R^2	57.5%			
Intercept (content)	0.0714503	0.0022332	31.99	$< 2e-16$
surprisal	0.0014206	0.0001441	9.859	$< 2.3e-13$
probMinus1	-0.0017461	0.0038742	-0.451	**0.65**
Intercept (function)	-0.0239773	0.0031336	-7.652	$< 2.7e-14$
length (content)	1.707e-02	2.499e-04	68.321	$< 2e-16$
length (function)	1.579e-03	4.627e-04	3.411	< 0.001
surprisal:probMinus1	-5.244e-04	3.561e-04	-1.473	**0.14**
s(logFreq)				$< 2e-16$
R^2	78.4%			

Table 7
GAM coefficients fitting BERT+fine-tuning fixation FPD in training (**top**) and test (**bottom**) data: FPD $\sim$ surprisal $\times$ probMinus1 + POSgroup $\times$ wordlength + s(logFreq).

	BERT FPD			
parametric coeff.	estimate	std. error	t value	pr(>|t|)
Intercept (content)	9.626e-02	4.765e-04	202.020	$< 2e-16$
surprisal	1.319e-03	3.027e-05	43.586	$< 2e-16$
probMinus1	-4.998e-03	8.245e-04	-6.0616	$< 1.3e-09$
Intercept (function)	-2.293e-02	6.937e-04	-33.053	$< 2e-16$
length (content)	1.019e-02	8.616e-05	118.232	$< 2e-16$
length (function)	2.892e-03	1.693e-04	17.0848	$< 2e-16$
surprisal:probMinus1	-3.874e-04	1.077e-04	-3.599	< 0.001
s(logFreq)				$< 2e-16$
R^2	75.6%			
Intercept (content)	0.0960782	0.0021829	44.014	$< 2e-16$
surprisal	0.0012786	0.0001409	9.073	$< 2.3e-13$
probMinus1	-0.0013508	0.0037907	-0.356	**0.72**
Intercept (function)	-0.0192904	0.0030629	-6.298	$< 3.4e-10$
length (content)	0.0102735	0.0003941	26.069	$< 2e-16$
length (function)	0.0027876	0.0007299	3.819	< 0.001
surprisal:probMinus1	-0.0008111	0.0004600	-1.763	**0.08**
s(logFreq)				$< 2e-16$
R^2	73.5%			

Table 8
GAM coefficients fitting BERT fixation FPD for the training (**top**) and test (**bottom**) settings: FPD $\sim$ surprisal $\times$ probMinus1 + POSgroup $\times$ wordlength + s(logFreq).

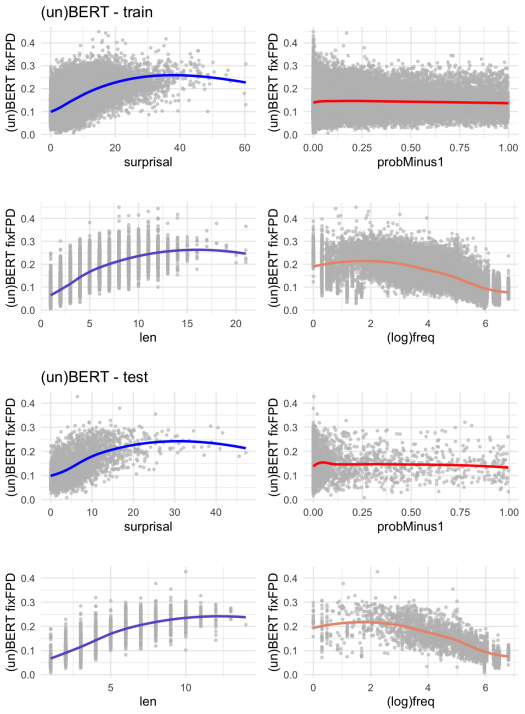

Figure 7: untuned BERT effects in training (**top panel**) and test (**bottom panel**) data, with surprisal, probability of the preceding token (*probMinus1*), word length (*len*) as predictors, word log-frequency as a smooth term (*logFreq*), and fixation first-pass duration as response variable.