# Using LSTM-based Language Models and human Eye Movements metrics to understand next-word predictions

Alfredo Umfurer[1], Juan Kamienkowski[2,3], and Bruno Bianchi[1,2,*]

[1] Departamento de Computación - FCEyN - UBA
[2] Instituto de Ciencias de la Computacion - CONICET
[3] Maestría en Exploración de Datos y Descubrimiento del Conocimiento - FCEyN - UBA
bbianchi@dc.uba.ar

**Abstract.** Modern Natural Language Processing (NLP) models can achieve great results resolving different types of linguistic tasks. This is possible thanks to a high volume of internal parametersthat are optimized during the training phase. They allow to model high-level linguistic properties. For example, LSTM-based language models have the ability to find long-term dependencies between words on a text, and use them to make predictions about upcoming words. Nevertheless, their complexity makes it hard to understand which features they use to generate predictions.
The neurolinguistic field faces a similar issue when studying how our brain processes language. For example, every adult reader has the ability to understand long texts and to make predictions of upcoming words. Nevertheless, our understanding on how these predictions are driven is limited. During the last decades, the study of eye movements during reading have shed some light on this topic, finding a relation between the time spent on a word (gaze duration) and its processing cost.
Here, we aim to understand how LSTM-based models predict future words and these predictions relate with human predictions, fitting statistical models commonly used in the neurolinguistic field with gaze duration as the dependent variable. We found that an AWD-LSTM Language Model can partially model eye movements, with high overlap with both human-Predictability and lexical frequency. Interestingly, this last overlap is seen to depend on the training corpus, being lower when the model is fine-tuned with a corpus similar to the one used for testing.

**Keywords:** LSTM · Eye Movements · Linear Mixed Models.

## 1 Introduction

The Natural Language Processing (NLP) field has witnessed a rapid evolution during the last years. This evolution has allowed to achieve the resolution of a great number of computational-linguistics tasks. Part of the advances performed in the last decade was made by the use of Long Short-Term Memory models,

firstly introduced in 1997 [8] and popularised some years ago after beating several competitions [18, 6, 17]. The main advantage of this type of Recurrent Neural Network (RNN) is the possibility to retain context information through long sequences of words. Such property is achieved by the use of forget gates, that solves the problem of the vanishing gradient.

Nevertheless, the advantage of using these complex models turns to be an issue when trying to understand how these networks make predictions and which mechanisms they use. For example, the high number of internal parameters that are optimized during the training phase makes it impossible to perform feature importance analyses.

In the Psycholinguistic field brain mechanisms involved in natural read are studied by relating word properties with behavioural and physiological data acquired from readers. For example, the Eye Tracking technique is based on recording the position of the reader's eyes on a screen during text presentation [13, 20, 15]. With this information, the time expended by the reader's eyes on each word (i.e., Gaze Duration –GD–) is analysed as a reflection of their processing cost. This variable is known to correlate with word properties like word length, lexical frequency, position in the sentence or text, and Predictability, among others [15, 14, 3]. Nowadays, these analyses are performed using Linear Mixed Models (LMM), that allows to understand how all these word properties relate with GD taking into account the variance introduced by subjects or the selected material for the experiment (random effects). Thus, by doing this type of analyses it is possible to understand which text features are used by our brains to process information.

Most of the those variables that correlates with GD can be easily estimated from the text or from an independent corpus. But the Predictability, that is defined as the probability of knowing a word before reading it, is a subjective variable. It is usually assessed by performing an experiment (named cloze-task) where a lot of participants are asked to answer the most probable word given an incomplete context [21]. As a consequence, Predictability is a hard and expensive variable to estimate.

Researchers had made several attempts to model it using simple computational models but, until now, they had not reached conclusive results [19, 9, 3, 10, 1]. In 2008, Ong and Kliegl [19] analysed how the conditional co-occurence probability (CCP) of a word given its context, measured by their frequency on internet search engines (Google, Yahoo!, MSN), and replaced the cloze-Predictability in Eye Movements models. They found that CCP acts like lexical frequency in predicting fixation durations. More recently, Hofmann and colleagues [9, 10] used NLP algorithms for next-word predictions. In these studies they trained N-grams, Recurrent Neural Networks, and Topic Models (LDA) with Wikipedia and movie subtitles, adding the resulting probabilities to statistical models with Eye Movement and electrophysiological variables as dependent variables. After analysing how much variance these probabilities account for in each model they conclude that computational algorithms can explain these human-based variables better than the original cloze-Predictability. But, they did not anal-

ysed how good these computer-based-Predictabilities are in replacing the cloze-Predictability. Finally, Algan [1] showed how a LSTM-based Predictability correlates with cloze-Predictability in Turkish.

In 2020 Bianchi and colleagues [3] showed that N-gram probabilities and semantic similarities from different distributional semantics algorithms (LSA, word2vec, FastText) can partially replace the cloze-Predictability on Linear Mixed Models (LMM) using the GD as the dependent Variable. In this study they analysed how much variance was left for the cloze-Predictability to explain the GD on the residuals of the LMM fitted with computational-Predictabilities. As far as we know, this is the only precedent of this type of analysis performed in Spanish.

Modelling Predictability will not only ease the experiments, but will also allow a better understanding of the human brain predictions. More importantly, nowadays we face the opportunity to understand both brain and RNN predictions by exploring their relationship. And fortunately, Psycholinguistics tools can help with this task.

In the present work, we aim to implement the AWD-LSTM model [17] to comprehend how predictions are performed, by analysing how they relate with GD and other word properties. In order to compare with previous results, we use an available corpus of short stories with Gaze Duration and Predictability measured for each word.

## 2 Methods

### 2.1 Eye movements

Eye movements were recorded from thirty-six native Spanish readers with normal or correct-to-normal vision. Participants read eight stories from the *Buenos Aires Corpus* [14] and gaze position was recorded with a video-based eye tracker (EyeLink 1000 from SR research). This data is publicly available from Bianchi et al. [3]. Then, gaze position was used to calculate the First Pass Reading Time or Gaze Duration (GD) on each word. GD is defined as the total time spent on a word before leaving it for the first time, i.e., the addition of all fixations in a word during the first pass, without counting future refixations. This eye movement variable will be use as the dependent variable in the statistic models used in this study.

### 2.2 Cloze Task

The cloze task is performed by presenting uncompleted texts to participants that have to answer the next most probable word for that context. The corpus from Bianchi et al. comprises cloze-Predictability (i.e., the probability of correctly guessing each word in a cloze task) from more than 1000 participants ($16 \pm 8$ per words) collected online [3]. It was performed using a custom-made web page where participants logged-in to find one of the eight selected stories randomly

assigned. After finishing a story participants were allowed to close the experiment or to continue with a new randomly assigned text. Participants that closed the experiment could return to following stories at any moment. This data is publicly available from Bianchi et al. [3].

### 2.3 AWD-LSTM predictability

The AWD-LSTM model trained to get the next-word probability (from now LSTM-Predictability) consisted of three stacked LSTMs layers with 400, 1152 and 400 dimensions respectively, and multiple dropout layers, as described in [17]. Both in the input and the output layers, each word was represented as a 400 dimensions embedding. This model was trained on a corpus taken from Spanish Wikipedia, with a total of 444,571 files and 2,751,415 tokens. It was then fine-tuned using a small corpus of 2,081 Spanish books and 535,068 tokens [3]. Results from the trained-only and the trained-and-fine-tuned models will be analysed separately. Both models were trained on 10 epochs, using a One-Cycle policy with a maximum learning rate of 0.002. The fine-tuning phase was performed in two steps. Firstly, only the encoder layer was tuned, using two epochs (*max learning rate = 0.026*). Secondly, all the parameters were tuned, using eight epochs (*max learning rate = 0.0026*). These models were used to perform next-word prediction for each word from the story corpus used on the Cloze-Task experiment previously presented. This data is publicly available at http://reading.liaa.dc.uba.ar/.

### 2.4 Statistical analysis

The logit of Predictability measures (both cloze and computational) will be used as co-variables in successive Linear Mixed Models (LMM) with Log-transformed Gaze Duration (GD) as dependent variable. LMMs also include a set of previously described co-variables (launch position, the inverse of the word length, the logarithm of lexical frequency, their interaction, and the position in the line, the text, and the sentence). Subject and text identifiers, and the fixated word as string, were used as random effects. Text variables and the results for the Ngram model, was publicly available by Bianchi et al. [3]. Linear Mixed Models have no need of hyperparameters to set or optimise [2].

The outputs of LMM are the estimates of the slopes and their errors (SD) for each of the fitted fixed factors. Then, t-values are calculated as the ratio between each slope and its SD. These values represent how far away from zero the slopes are. As our models are fitted with a high number of instances, the distributions of used co-variables can be considered as normal, and thus, absolute t-values larger than 2.0 are considered significant with $\alpha < 0.05$ [4]. Each significant effect implies a linear relation between that covariable and the dependent variable. Since the estimate of the slope of a LMM co-variable depend on its scale, and each estimation of Predictability has a different range of values, we based our analyses mostly on the effect significance and t-values, which are standardized.

To analyse how each computer-Predictability mimics the cloze-Predictability, residuals of the corresponding LMM will be analysed. That is, after fitting a LMM with a computer-Predictability as co-variable, residuals of the fixed effects will be used in a new LMM with cloze-Predictability as the only fixed effect, conserving the random structure. For this procedure, we used the *remef* function [11] implementation for R.

We used the Akaike Information Criterion (AIC) to compare between different hierarchically built models on the same data. This estimator is calculated as the log likelihood of the model, compensated by the number of fixed effects [24]. The smaller it gets, the better the model to explain the data, compensating the number of variables to avoid overfitting.

## 3 Results

A series of LMMs with different combinations of co-variables were fitted to analyse how the AWD-LSTM model mimic the cloze-Predictability. The baseline model (Fig. 1A, M0) comprised a set of previously described co-variables: launch position, the inverse word length, the log lexical frequency, and their interaction, and the position in the line, the text and the sentence. They all showed significant effects as expected from previous studies [14, 3]. Subsequently, the cloze-Predictability was added in a subsequent model (Fig. 1A, M1), showing a clear negative effect on GD. The addition of this co-variable generated negligible changes on the co-variables effects of the baseline model.

Results from two AWD-LSTM models were added as co-variables in independent LMMs (LSTM-Predictability). Firstly, we used the output of a LSTM model trained only with a Spanish Wikipedia corpus, a big but no specific corpus for the task (Fig. 1A, M2). The t-value of Wikipedia-Only LSTM-Predictability on the LMM ($t = -14.97$) was similar to the cloze-Predictability ($t = -16.23$). Additionally, some co-variables from the baseline model showed changes on their effects, particularly the lexical frequency.

Going one step further, we can compare not only the amount of variance explained by the covariables, but also if it is the same portion of it. Then, to observe if the cloze-Predictability can be explained by the results from this LSTM, the residuals of the LMM (M2 Wiki-Only) were fitted in a new LMM with cloze-Predictability as the only fixed effect (Residuals + cloze-Predictability). This analysis showed that the effect of human Predictability remains significant ($t = -11.63$, Fig. 1B, M2). This implies there is still variance associated with cloze-Predictability left after fitting the model with the LSTM results. That is, AWD-LSTM trained with a Wikipedia corpus can only partially model the cloze-Predictability effect on Gaze Duration. Moreover, the drop in the frequency effect significance shows that a part of its effect comes from lexical frequency.

Secondly, the output of an AWD-LSTM model trained with Spanish Wikipedia and fine-tuned with a corpus of stories was included as a co-variable (Fig. 1A, M3). Its t-value on the LMM was almost the same as cloze-Predictability ($t = -16.76$). Contrary to the observed result for the M2, the frequency ef-

fect remained significant, although largely decreased. Furthermore, the cloze-Predictability effect on the residuals of the LMM in M3 was smaller than in M2 ($t = -9.87$, Fig. 1B, M3), suggesting that the fine-tuning improved the LSTM performance.

These effects also have an impact on the goodness-of-fit of the fitted LMMs, estimated with the Akaike Information Criterion (AIC) for each model relative to M0 and M1 (Fig. 1C). In particular, M2 showed an increase on the absolute AIC relative to M0 and a decrease relative to M1. This indicates a better fit than the baseline model, but worst than the Cloze Model (M1). Meanwhile, M3 showed a slight improvement on the overall fitting relative to M1.

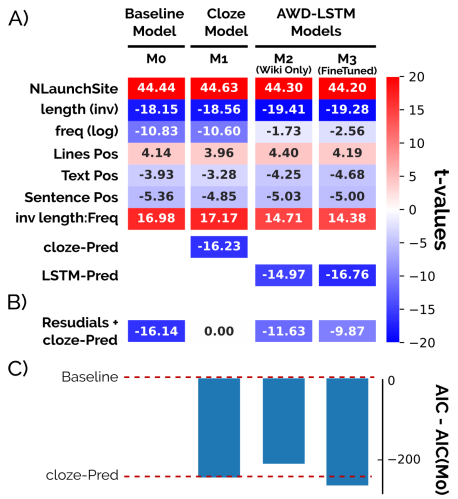


**Fig. 1. A)** t-values from 4 LMMs with different set of co-variables. **M0**: baseline models. **M1**: baseline model and Cloze-Predictability variable. **M2**: baseline model and LSTM-Predictability trained only with Wikipedia. **M3**: baseline model and LSTM-Predictability trained with Wikipedia and fine-tuned with a story corpus. **B)** t-values for the cloze-Predictability effect on a Linear Mixed Model fitted on the residuals of each of models on A. **C)** AIC values for each of the fitted models on A relative to the M0 AIC.

Bianchi et al. [3] explored the output of a 4-gram model as co-variable in the same corpus, showing significant effects on the LMM, also with a decrease in the frequency effect (Fig. 2A, M4). The addition of the LSTM-Predictability from M3 (fine-tuned) also had an impact on the LMM (Fig. 2A, M5), measured both in the co-variable ($t = 6.11$) and in the increase of the absolute AIC value (Fig. 2C). Finally, there was a significant effect of cloze-Predictability when fitting the residuals of the LMMs (Fig. 2B, M5). This suggests that the effect of cloze-Predictability on Gaze Duration cannot be fully explained by these computational models.

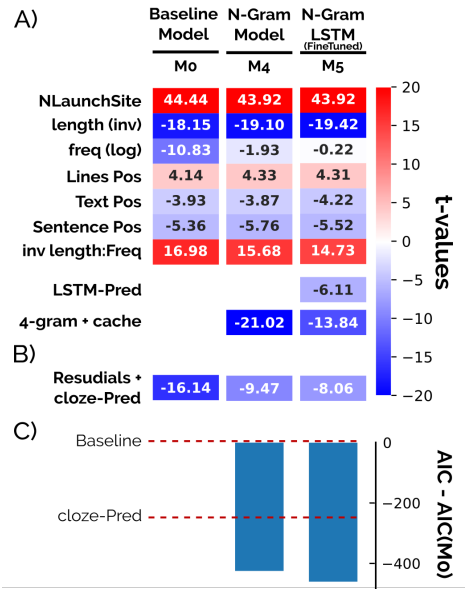


**Fig. 2. A)** t-values from 2 more LMMs with different set of co-variables. **M0**: baseline models. **M4**: baseline model and Ngram model from [3], **M5**: baseline model, Ngram model, and LSTM-Predictability from the fine-tuned model. **B)** t-values for the cloze-Predictability effect on a Linear Mixed Model fitted on the residuals of each of models on A. **C)** AIC values for each of the fitted models on A relative to the M0 AIC.

## 4 Conclusions

LSTM networks have allowed great advances in Natural Language Processing tasks. Their large number of internal parameters and their internal architecture that avoids the problem of the vanishing and exploding gradients, allow them to learn complex interactions while keeping context information. But, at the same time, it generates highly opaque models when trying to understand what features are used to abstract the language. In the present work we explored how a LSTM network models natural language using its output to mimic a human-based linguistic variable. Cloze-Predictability is a commonly used variable in psycholinguistic research when studying how our brain process language. This variable is known to correlates with behavioural (e.g. Fixation Duration) [15] and electrophysiological metrics (e.g. scalp potentials) [16]. For this study we replaced it with the LSTM-Predictability on Linear Mixed Models (LMM), statistical models that are used to understand brain processes.

Using a fraction of the Spanish Wikipedia, we trained a LSTM model (trained-only). This model was then fine-tuned with a small corpus of narrative texts. Both models were used to estimate LSTM-Predictabilities on a set of 8 stories, previously used by Bianchi et al. [3] for a similar analysis. LSTM-Predictabilities were used as co-variables in independent LMM with other linguistic proper-

ties and the Gaze Duration as dependent variable. Predictabilities from both the trained-only and the fine-tuned models showed significant effects, overlap with the cloze-Predictability, and an improvement on the LMM goodness of fit (measured with the AIC) relative to baseline. But, on the trained-only model there was a loss of the lexical frequency effect, becoming non-significant. This highlights that, on the one side, to achieve a good replacement of the cloze-Predictability it is important to consider training or fine-tuning the computational model on a corpus similar to the tested one. A general corpus, like Wikipedia, will lead to perform predictions based mostly on word Frequency. On the other side, this shows that a computational-Predictability generating a better goodness of fit than the original cloze-Predictability does not imply that the former one is better for explaining brain processes underlying predictions.

In this line, future work must be aimed to improve the LSTM-Predictability based on AWD-LSTM and other LSTM architectures, experimenting with different parameters on the training and testing phases. First, using a larger corpus for the specific fine-tuning may result in a better replacement of the cloze-Predictability, allowing to further explore how LSTM predictions are performed. Secondly, experimenting with the amount of information used by the LSTM to predict future words would give more insight on how long dependencies are used by the model and, also, by the brain. Moreover, these analyses could be extended to other more modern models, like transformers based models. In order to do so, it is important to take into account the difficulty and cost of their training.

Finally, aside of how good AWD-LSTM could be on predicting upcoming words, it only partially explained the effect of Predictability on the cognitive processing of words. Moreover, it had a large impact on the frequency effect, something that is not present on the classical cloze-Predictability effect. Thus, to predict future words, LSTM seems to rely on the lexical frequency more than humans. This overlap with the frequency effect was previously observed for a Ngram model. The comparison between the AWD-LSTM model presented here and the Ngram model implemented by Bianchi and collaborators [3] showed that they explained different aspects of the cloze-Predictability, with some degree of overlapping. Thus, the comparison with simpler and more transparent models may also serve as a way to understand complex models, like LSTMs. Despite the N-gram model can be improved, for example adding information about grammatical properties of words [5], the text processing needed for this (like Part-of-Speech tagging) is highly expensive and not robust, while modern NLP algorithms, like AWD-LSTM, can infer this information implicitly. Additionally, algorithms based on neural networks have more hyperparameters (embedding size, number of layers, etc.) that were not explored in the present study and may allow future improvements.

This work is another step in the dialogue between NLP and Neuroscience, using cognitive and physiological measures to understand NLP and vice versa, that will boost both fields [23, 7, 12, 22].

## Acknowledgements

The authors were supported by the Consejo Nacional de Investigaciones Cientificas y Técnicas (CONICET), the Universidad de Buenos Aires (UBA), and the Programa de Pasantías from the Departamento de Ciencias de la Computación (Facultad de Ciencias Exactas y Naturales, UBA). The research was supported by the UBA (20020190100054BA), the National Agency of Promotion of Science and Technology (PICT 2018-2699) and the CONICET (PIP 11220150100787CO).

## References

1. Algan, A.C.: Prediction of words in Turkish sentences by LSTM-based language modeling. Master's thesis, Middle East Technical University (2021)
2. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823 (2014)
3. Bianchi, B., Monzón, G.B., Ferrer, L., Slezak, D.F., Shalom, D.E., Kamienkowski, J.E.: Human and computer estimations of predictability of words in written language. Scientific reports **10**(1), 1–11 (2020). https://doi.org/https://doi.org/10.1038/s41598-020-61353-z
4. Bianchi, B., Shalom, D.E., Kamienkowski, J.E.: Predicting known sentences: Neural basis of proverb reading using non-parametric statistical testing and mixed-effects models. Frontiers in human neuroscience **13**, 82 (2019). https://doi.org/https://doi.org/10.3389/fnhum.2019.00082
5. Bilmes, J., Kirchhoff, K.: Factored language models and generalized parallel backoff. In: Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers. pp. 4–6 (2003)
6. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp. 6645–6649. IEEE (2013). https://doi.org/https://doi.org/10.1109/ICASSP.2013.6638947
7. Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-inspired artificial intelligence. Neuron **95**(2), 245–258 (2017)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
9. Hofmann, M.J., Biemann, C., Remus, S.: Benchmarking n-grams, topic models and recurrent neural networks by cloze completions, eegs and eye movements. In: Cognitive approach to natural language processing, pp. 197–215. Elsevier (2017). https://doi.org/https://doi.org/10.1016/B978-1-78548-253-3.50010-X
10. Hofmann, M.J., Remus, S., Biemann, C., Radach, R.: Language models explain word reading times better than empirical predictability. PsyArXiv (2020). https://doi.org/https://doi.org/10.31234/osf.io/u43p7
11. Hohenstein, S., Kliegl, R.: remef (remove effects)(version v0. 6.10) (2013)
12. Hollenstein, N., Torre, A., Zhang, C.: Cognival: Framework for cognitive word embedding evaluation. arXiv:1909.09001 (2019)
13. Just, M.A., Carpenter, P.A.: A theory of reading: From eye fixations to comprehension. Psychological review **87**(4), 329 (1980). https://doi.org/https://psycnet.apa.org/doi/10.1037/0033-295X.87.4.329

14. Kamienkowski, J.E., Carbajal, M.J., Bianchi, B., Sigman, M., Shalom, D.E.: Cumulative repetition effects across multiple readings of a word: Evidence from eye movements. Discourse Processes **55**(3), 256–271 (2018). https://doi.org/https://doi.org/10.1080/0163853X.2016.1234872
15. Kliegl, R., Nuthmann, A., Engbert, R.: Tracking the mind during reading: The influence of past, present, and future words on fixation durations. Journal of experimental psychology: General **135**(1), 12 (2006). https://doi.org/https://psycnet.apa.org/doi/10.1037/0096-3445.135.1.12
16. Kutas, M., Hillyard, S.A.: Reading senseless sentences: Brain potentials reflect semantic incongruity. Science **207**(4427), 203–205 (1980). https://doi.org/10.1126/science.7350657
17. Merity, S., Keskar, N.S., Socher, R.: Regularizing and optimizing lstm language models. arXiv preprint arXiv:1708.02182 (2017). https://doi.org/https://arxiv.org/abs/1708.02182v1
18. Märgner, V., Abed, H.E.: Icdar 2009 arabic handwriting recognition competition. In: 2009 10th International Conference on Document Analysis and Recognition. pp. 1383–1387 (2009). https://doi.org/https://doi.org/10.1109/ICDAR.2009.256
19. Ong, J.K., Kliegl, R.: Conditional co-occurrence probability acts like frequency in predicting fixation durations. Journal of Eye Movement Research **2**(1) (2008). https://doi.org/https://doi.org/10.16910/jemr.2.1.3
20. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological bulletin **124**(3), 372 (1998). https://doi.org/https://psycnet.apa.org/doi/10.1037/0033-2909.124.3.372
21. Taylor, W.L.: "cloze procedure": A new tool for measuring readability. Journalism quarterly **30**(4), 415–433 (1953). https://doi.org/https://doi.org/10.11772F107769905303000401
22. Toneva, M., Wehbe, L.: Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In: Advances in Neural Information Processing Systems. pp. 14954–14964 (2019)
23. Ullman, S.: Using neuroscience to develop artificial intelligence. Science **363**(6428), 692–693 (2019)
24. Vaida, F., Blanchard, S.: Conditional akaike information for mixed-effects models. Biometrika **92**(2), 351–370 (2005)