Exploring the Integration of Eye Movement Data on Word Embeddings

Fermín Travi[1,2], Gabriel Leclercq[2], Diego Fernández Slezak[1,2], Bruno Bianchi[1,2], Juan E. Kamienkowski[1,2,3]

[1]Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

[2]Instituto de Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

[3]Maestría en Explotación de Datos y Descubrimiento del Conocimiento, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

Correspondence: ftravi@dc.uba.ar

Abstract

Reading, while structured, is a non-linear process. Readers may skip some words, linger on others, or revisit earlier text. Emerging work has started exploring the incorporation of reading behaviour through eye-tracking into the training of specific language tasks. In this work, we investigate the broader question of how gaze data can shape word embeddings by using text as read by human participants and predicting gaze measures from them. To that end, we conducted an eye-tracking experiment with 76 participants reading 20 short stories in Spanish and fine-tuned Word2Vec and LSTM models on the collected data. Evaluations with representational similarity analysis and word pair similarities showed a limited, but largely consistent, gain from gaze incorporation, suggesting future work should expand linguistic diversity and use cognitively aligned evaluations to better understand its role in bridging computational and human language representations.

1 Introduction

The field of natural language processing (NLP) is currently driven by artificial neural networks primarily trained on the task of predicting the next word in a given sentence (Radford et al., 2018, 2019). However, next-word prediction of written text is a pale reflection of how language is processed in the brain, as written text is the product of deliberate conscious processes, often edited, proofread, and restructured. This stands in stark contrast with the spontaneous generation of language in our everyday life.

Eye-tracking during reading has long been recognized as a central tool for unraveling language processing in the brain (Kliegl et al., 2006), with its earliest studies dating over a century ago (Huey, 1908). Recent advances have demonstrated how eye movement data from reading can be integrated into NLP applications, enhancing performance in various downstream language tasks through its incorporation into language models (Zhang and Hollenstein, 2024; Yang and Hollenstein, 2023; Hollenstein and Zhang, 2019; Klerke et al., 2016; Barrett et al., 2018). However, whether gaze information can influence the latent representations of language processing models to align more closely with human cognitive processing remains an unresolved question. Moreover, despite its long history in reading research, eye-tracking during reading datasets are not widely available, and most of this emerging research has focused exclusively on English, while similar studies utilizing eye-tracking data in Spanish have yet to be conducted.

Thus, we collected eye-tracking data during reading in Spanish and utilized the resulting gaze information to train simple language processing models. Our findings suggest that incorporating gaze information into word embeddings may offer modest but promising steps toward greater cognitive alignment. We argue that combining larger, linguistically diverse datasets with cognitively focused evaluation tasks will be critical for uncovering the potential of eye-tracking to bridge computational and human language representations.

2 Materials & Methods

2.1 Eye-tracking experiment

To collect eye movement patterns during natural reading, we selected twenty self-contained short stories (800 ($\pm$ 135) words long, average reading time of three minutes) written in Latin American Spanish. We cleaned and processed data from 76 participants (mean age 23.5 (IQR 4.8); 44 females, 32 males; mostly college students), resulting in 1,015 trials (Fig. A1). All details of the experiment can be found in Appendix A. Gaze measures were extracted from those words that were not the first or last words in a sentence or line and did not contain punctuation marks, dashes or numbers, resulting in

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 55–65

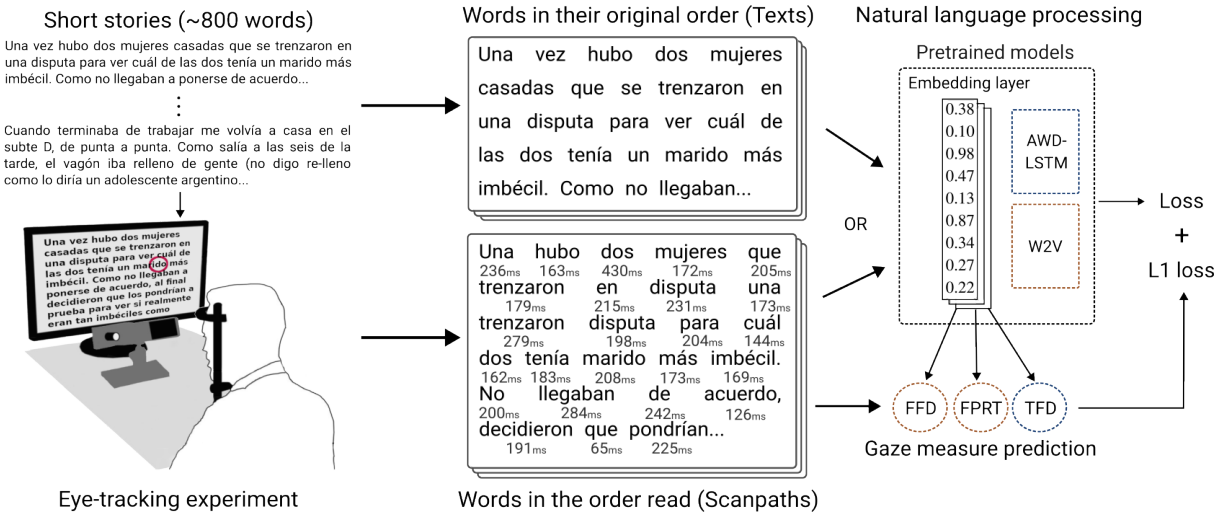

Figure 1: Gaze embedding pipeline. The stories read during the eye-tracking experiment were reconstructed following the reading order of the participants (Scanpaths). Gaze measures were extracted from all trials, discretized in ten bins for each individual, and a global average for each word was computed. These values were then predicted from the word embeddings as the output of a fully connected layer.

3,016 unique words (of 3,493) with gaze measures. The code used for the experiment and extracting gaze measures can be found at https://github.com/NeuroLIAA/reading-et.

2.2 Natural Language Processing

To test our hypothesis, we selected two different architectures: a word embedding model (Word2Vec in its skip-gram variation with negative sampling (Mikolov et al., 2013)), and a language model (AWD-LSTM (Merity et al., 2017), one of the latest variations of LSTMs), from which we extracted its embedding layer. These selections were based primarily on simplicity and ability to perform well without requiring extensive amounts of data. While Large Language Models could be applied with larger gaze datasets, these simpler models serve as an efficient proof of concept for our methodology. The hardware employed for training consisted of a personal computer: Intel i7-11700, 32GB RAM DDR4, and a GPU ASUS RTX 3060 12GB. The pre-training and fine-tuning consumed, respectively, 5 h and 0.5 h for the Word2Vec model and 50 h and 2 h for the LSTM model. The source code is available at https://github.com/NeuroLIAA/gaze-word-embeddings.

2.2.1 Baselines

As baseline training data, we employed a 2019 dump of Spanish Wikis hosted in Huggingface[1]. To reduce vocabulary size and noise, words that contained numbers, special characters, non-latin characters, were shorter than two tokens, or longer than twenty tokens, were filtered out. Resulting sentences shorter than four words and longer than forty were also removed. This yielded a total of approximately 22 million sentences. Vocabulary was composed of individual words that appeared at least twenty times. Pre-training for both Word2Vec and AWD-LSTM followed default hyperparameters (with an embedding size of 300 and batch size of 32 sentences) for five epochs.

2.2.2 Gaze embedding

To embed gaze, we used scanpath-generated text as input (extracted from the eye-tracking experiment described in §2.1) and fine-tuned the baseline with them, while also predicting gaze measures from word embeddings (Fig. 1). Scanpath-generated text was constructed by following the fixations from each trial in the experiment, yielding 1,015 different texts (named *Scanpaths*). Although it mostly overlapped with the text read, its word order was markedly distinct from written text, as human reading is a non-sequential process. If a word with a punctuation mark was fixed several times consecutively, or a regression was done inside the same sentence, all those words were stripped from the punctuation marks and put together as a single sentence. This totalled 44,748 sentences (612,299 words) after preprocessing. For a correct comparison, we also defined a corpus containing the original texts (from §2.1) repeated the same number of

[1]https://huggingface.co/datasets/large_spanish_corpus

		CKA	Correlations		
			SimLex	Abstract	Concrete
W2V	Baseline	**0.1434**	**0.4147**	**0.4814**	**0.3312**
	Scanpaths	0.1402*	0.2946*	0.3357*	0.3036*
	Scanpaths + GM	**0.1434**	0.3867*	0.3163*	0.3308
	Texts	0.1382*	0.3119*	0.2796*	0.2671*
	Texts + GM	0.1422*	0.4018*	0.2784*	0.3006*
AWD-LSTM	Baseline	0.1114	0.2301	0.2507	0.1238
	Scanpaths	0.1088*	0.2377*	0.2147*	**0.2199***
	Scanpaths + GM	0.1102*	**0.3537***	0.2261*	0.2122*
	Texts	0.1124*	0.3113*	**0.2715***	0.1741*
	Texts + GM	**0.1127***	0.3298*	0.2689*	0.1773*

Table 1: Mean CKA to word embeddings derived from SWOW-RP of 1,650 content words with gaze measures, and mean Spearman rank correlation of cosine distances between word pairs with human similarity judgments. The latter analysis was conducted across three datasets (SimLex, Abstract, and Concrete) with 216, 276, and 378 word pairs respectively. Random samplings of 1000 words and 100 word pairs were performed a hundred times with replacement. Baseline refers to models trained on the baseline corpus, Scanpaths are models fine-tuned on text as read by participants, and Texts are models fine-tuned on stimuli as is. GM refers to the addition of gaze measures to the training process. The asterisk indicates the distribution was significantly different from the baseline.

times, totalling 42,213 sentences (666,374 words, named *Texts*).

Gaze measures are usually classed as early (First Fixation Duration, FFD; First Pass Reading Time, FPRT) or late (Total Fixation Duration, TFD), depending on the stage of reading processing they reflect (Inhoff, 1984). Early measures are thought to be a reflection of primarily automatic word recognition and lexical access processes, whereas late measures tend to reflect more conscious, controlled, strategic processes. By forcing the word model to predict them (early measures for Word2Vec, late for AWD-LSTM), we intended to embed (cognitive) attention into it (Klerke et al., 2016; Barrett et al., 2018). Specifically, this was done by adding a fully connected layer that received the embedding of the input word and predicted its corresponding gaze measures. The resulting L1 loss was then added to the standard loss of the model (Fig. 1).

These measures were computed individually for each item and participant in the experiment. As the values of these measures vary between participants and items, we discretized them in ten quantiles for each participant. A word average was then obtained by computing the mean across participants and items. If the input word contained no gaze measure, its value was set to zero. It is important to note that these measures are not independent, as First Fixation Duration (FFD) is a part of First Pass Reading Time (FPRT), which is a part of Total Fixation Duration (TFD). We left out regression measures because they showed to be lowly correlated between subjects (Fig. A2). Fine-tuning was carried out for 50 epochs with the same hyperparameters as the baseline.

2.3 Evaluation framework

2.3.1 Association-based word embeddings

Our main reference point was based on a massive word association task in Rioplatense Spanish (*Small World of Words Rioplatense Spanish* (SWOW-RP)) (Cabana et al., 2024), primarily due to its size, as well as the well-studied link between word associations and semantic representations stored in memory (De Deyne et al., 2016). From these word associations, Cabana et al. (2024) derived graph embeddings that have been shown to be more closely related to human similarity judgments of word pairs than standard word embeddings.

We hypothesize that gaze-derived word embeddings are a closer match to word embeddings derived from SWOW-RP with respect to the baseline. To evaluate this, we employed centered kernel alignment (CKA) (Kornblith et al., 2019) on the resulting 1,650 content words with gaze measures. CKA is a similarity measure that quantifies the similarity between two sets of representations by comparing their Gram matrices in a reproducing kernel Hilbert space. Unlike traditional similarity metrics, CKA is invariant to orthogonal transformations and can effectively capture global structural similarities between high-dimensional representa-

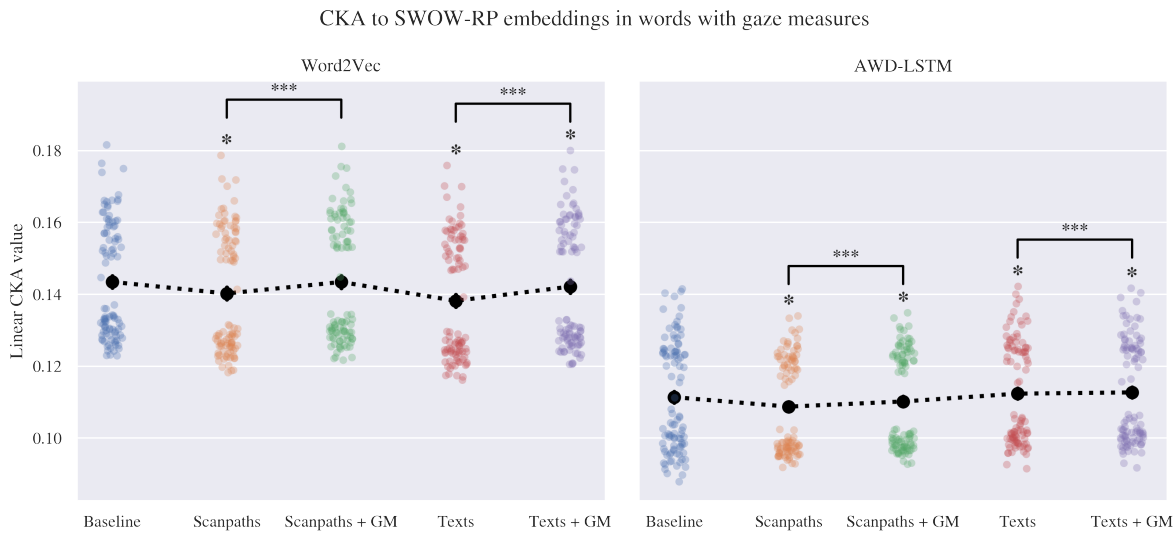

Figure 2: Distribution of the CKA values to the word embeddings derived from SWOW-RP presented in Table 1.

tions, making it particularly useful for comparing embedding spaces across different domains or models. Recent works in machine learning and cognitive science have leveraged CKA to analyze representational similarities in neural networks, comparing learned representations across different layers, architectures, and even modalities (Vulić et al., 2020b; Hao et al., 2023; Maniparambil et al., 2024). In the context of our study, CKA provided a robust method to assess the semantic alignment between gaze-derived and association-based word embeddings, allowing us to evaluate how closely these distinct representational spaces match. CKA similarity ranges between 0 and 1, from most dissimilar to most similar spaces. To mitigate potential outlier effects, we performed random samplings of 1,000 words with replacement a hundred times and report the mean and standard error to it.

2.3.2 Word pairs similarity judgments

A more classical way of evaluating word embeddings is to compute the Spearman rank correlation between the cosine distance of two words and their corresponding semantic similarity as defined by human participants (Mikolov et al., 2013). However, a limitation to this approach is the requirement for both words to have been fine-tuned. We made use of two different resources: Multi-SimLex ES (*SimLex*; 1,888 semantically aligned concept pairs, of which 216 possess gaze measures) (Vulić et al., 2020a) and a relatedness task for Rioplatense Spanish speakers collected by De Deyne et al. (2020) (3,321 conceptually abstract (*Abstract*) and 3,321 conceptually concrete (*Concrete*) word pairs, of which 276 and 378 possess gaze measures, respectively). In this case, we sampled 100 word pairs randomly with replacement a hundred times.

3 Results

We evaluated the impact of fine-tuning NLP models using text as read by participants in an eye-tracking experiment, compared to text in its original order, as well as the effect of predicting gaze measures from word embeddings during training. Both architectures successfully incorporated gaze measures into their embeddings: Word2Vec achieved a close-to-perfect correlation between predicted and true gaze values, while AWD-LSTM achieved a median correlation of 0.89 per batch. All reported p-values were computed using Wilcoxon signed-rank tests on bootstrapped distributions.

When comparing the alignment of these embeddings with those derived from SWOW-RP, differences emerged (Fig. 2). For Word2Vec, fine-tuning without gaze measures resulted in a slight decrease in alignment compared to the baseline (Table 1). However, the incorporation of gaze measures slightly improved their alignment and their distributions resulted significantly different from their counterparts without gaze measures (both $p < 0.0001$, ws. 0.0), although the addition of gaze measures to Scanpaths resulted in CKA values not significantly different to the baseline. In contrast, AWD-LSTM showed no decrease when fine-tuned with Texts relative to the baseline and a slight decrease when fine-tuned with Scanpaths. The addition of gaze measures barely increased

the mean, but provided significantly different distributions ($p < 0.0001$, ws. 1189.0 for Texts and $p < 0.01$, ws. 1698.0 for Scanpaths).

The differing behavior between architectures when fine-tuning with Scanpaths or Text likely stems from their design: Word2Vec, employing a bag-of-words approach with moving windows, is less affected by syntax and may leverage the non-sequential patterns of human reading when capturing first-order relationships. Meanwhile, AWD-LSTM, pre-trained on syntactically structured sentences, is more sensitive to deviations from natural text order, such as those found in Scanpaths.

When analyzing the correlation between the cosine distance of word embeddings and human similarity judgments of word pairs with gaze measures, distinct trends emerged for the two models tested (Fig. B1). For Word2Vec, fine-tuning generally reduced correlation scores compared to the baseline across datasets, with the most pronounced drop observed in Abstract (Table 1). All correlations were significantly different to the baseline ($p < 0.0005$), with the exception of Scanpaths plus gaze measures in Concrete ($p > 0.5$, ws. 2448.0), as was the case in the CKA analysis. The inclusion of gaze measures to Scanpaths and Texts increased correlations in all datasets except for Abstract, as well as producing significantly different distributions ($p < 0.0001$). The fine-tuned variations of AWD-LSTM, on the other hand, improved correlations with respect to the baseline, except for Abstract as well. Correlations were once again significantly different to the baseline ($p < 0.001$) and the addition of gaze measures had a significant impact in SimLex, but not in Abstract nor Concrete.

In the case of word pairs not present in the stimuli, the mean remained relatively unchanged across datasets for both models (Fig. B2).

4 Discussion

In this work, we investigated the integration of gaze information into word embeddings of language processing models by means of feeding them text as read by human participants and incorporating gaze measures into the latent space. Our findings suggest that architectures like Word2Vec may better leverage the non-sequential patterns of human reading, while pre-trained language models like AWD-LSTM appear to be more negatively impacted by them. Moreover, the incorporation of early and late gaze measures, respectively, yielded modest improvements in most evaluation tasks, hinting at the potential for gaze measures to nudge the latent space toward greater cognitive alignment. However, further research should look deeper into the morphological or linguistic variables involved to fully understand and optimize this effect. One key limitation of our study lies in some aspects of the dataset: while the number of participants in our eye-tracking experiment was high (76), the number of unique words read was low (3,493). This constraints the size of the fine-tuning corpus (44,709 sentences) and the contextual variety it offers. As seen in analogous studies translating other cognitive modalities to deep learning models (Tang et al., 2023), future efforts should prioritize increasing the number of reading sessions per subject rather than expanding the participant pool. Additionally, tasks more closely tied to cognitive processes, such as cloze tasks (Bianchi et al., 2020), may provide a more suitable evaluation framework.

Finally, our study is the first to integrate Spanish eye-tracking data into language models, raising questions about the language-specificity of prior findings. Future work can build on this foundation to enhance the use of gaze data in aligning computational and human language representations.

5 Limitations

As discussed throughout the article, the present study is limited by the contextual diversity and number of unique words present in the eye-tracking experiment. Vocabulary size is small, which, in turn, constrains the size of the evaluation space. Future work will expand this experiment by including novel texts. The lack of linguistic resources in Spanish also makes it impossible to combine datasets.

Gaze measures were aggregated to obtain global averages, but there may be large individual variability across participants (see Fig. A2). In line with recent suggestions in functional magnetic resonance images (fMRI) (Kupers et al., 2024), future work will include several sessions per participant. Eye-tracking offers the advantage of enabling intensive sampling across a substantial number of participants.

References

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of*

the 22nd Conference on Computational Natural Language Learning, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Bruno Bianchi, Gastón Bengolea Monzón, Luciana Ferrer, Diego Fernández Slezak, Diego E. Shalom, and Juan E. Kamienkowski. 2020. Human and computer estimations of predictability of words in written language. *Scientific Reports*, 10(1):4396. Publisher: Nature Publishing Group.

David H. Brainard. 1997. The Psychophysics Toolbox. *Spatial Vision*, 10(4):433–436. Place: Netherlands Publisher: VSP.

Álvaro Cabana, Camila Zugarramurdi, Juan C. Valle-Lisboa, and Simon De Deyne. 2024. The "Small World of Words" free association norms for Rioplatense Spanish. *Behavior Research Methods*, 56(2):968–985.

Hernán Casciari. 2021. *100 covers de cuentos clásicos*. Editorial Orsai S.R.L., Buenos Aires.

Simon De Deyne, Álvaro Cabana, Bing Li, Qiang Cai, and Matthew McKague. 2020. A cross-linguistic study into the contribution of affective connotation in the lexico-semantic representation of concrete and abstract concepts. In *Proceedings of the Annual Conference of the Cognitive Science Society*, volume 7.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1861–1870, Osaka, Japan. The COLING 2016 Organizing Committee.

Andrew Duchon, Manuel Perea, Nuria Sebastián-Gallés, Antonia Martí, and Manuel Carreiras. 2013. EsPal: One-stop shopping for spanish word properties. *Behavior Research Methods*, 45(4):1246–1258.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. 2023. One-for-all: bridge the gap between heterogeneous architectures in knowledge distillation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. 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 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

E B Huey. 1908. *The Psychology and Pedagogy of Reading*. The Psychology and Pedagogy of Reading. Macmillan, Oxford, England.

Albrecht Werner Inhoff. 1984. Two stages of word processing during eye fixations in the reading of prose. *Journal of Verbal Learning and Verbal Behavior*, 23(5):612–624.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Reinhold Kliegl, Antje Nuthmann, and Ralf Engbert. 2006. Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology. General*, 135(1):12–35.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. 2019. Similarity of neural network representations revisited. *CoRR*, abs/1905.00414.

Eline R. Kupers, Tomas Knapen, Elisha P. Merriam, and Kendrick N. Kay. 2024. Principles of intensive human neuroimaging. *Trends in Neurosciences*, 47(11):856–864.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Mohamed El Amine Seddik, Karttikeya Mangalam, and Noel E. O'Connor. 2024. Do vision and language encoders represent the world similarly? *Preprint*, arXiv:2401.05224.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing LSTM language models. *CoRR*, abs/1708.02182.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Núria Sebastián Gallés, Antonia Martí, and Manuel Carreiras. 1998. *LEXESP: Léxico informatizado del español*. Ediciones de la Universidad de Barcelona, Barcelona.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866. Publisher: Nature Publishing Group.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020a. Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4):847–897.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020b. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Duo Yang and Nora Hollenstein. 2023. PLM-AS: Pretrained Language Models Augmented with Scanpaths for Sentiment Classification. *Proceedings of the Northern Lights Deep Learning Workshop*, 4.

Leran Zhang and Nora Hollenstein. 2024. Eye-tracking features masking transformer attention in question-answering tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7057–7070, Torino, Italia. ELRA and ICCL.

A Eye-tracking experiment

All participants were native Spanish speakers and had normal or corrected-to-normal vision. All of them were recruited from the university mailing lists and were compensated with the equivalent of 5 USD per one-hour session. Written informed consent in agreement with the Helsinki declaration was provided by each of them. The experiment was approved by the Comité de Ética del Centro de Educación Médica e Investigaciones Clínicas "Norberto Quirno" (CEMIC) (Protocol 435). Records were anonymized in compliance with ethical board approvals and contain no personal information.

A.1 Corpus

Fifteen of the twenty short stories were extracted from "100 covers de cuentos clásicos" (Casciari, 2021), while the other five were extracted from online Argentinian blog posts. The original stories of the former were written by several different authors and were subsequently simplified, translated (if needed) and re-written in Spanish by Hernán Casciari. This way, there is diversity in literary style, while maintaining both difficulty and slang constant. The titles, authors and fixation statistics can be found in Table A1.

The selection criteria for the short stories was based on minimizing dialogue, very short and very long sentences (less than six words and greater than 29 words, respectively), infrequent words (less than 100 appearances in the Latin American subtitles database EsPal (Duchon et al., 2013)), infrequent characters (¿; ?; ¡; !; "; "; —; «; (;)), not containing written dates, and being no shorter than 400 words and no longer than 1500 words.

A.2 Environment & Setup

The experiment was written in MATLAB 2015a, using Psychtoolbox-3 (Brainard, 1997). It was conducted in a dark room, employing the EyeLink 1000 (SR Research, Ontario, Canada) binocular eye-tracker at 1000Hz. The monitor had a resolution of 1920x1080 and participants were seated 55cm away from it, using a chin and forehead rest to stabilize their head. The stimuli were presented in Courier New with font size 24 and black color with a gray background, 55 pixels of line spacing, 280 pixels of left margin and 185 pixels of top margin, with a maximum of fourteen lines per screen. Using these parameters, the text was divided in screens (ranging from four to six, depending on its length), and participants were allowed to go backward and forward between screens.

At the beginning of the experiment, participants were instructed to read the texts carefully, as comprehension questions would be asked at the end of each of them. Each short story constitutes an item and a separate trial, with eye-tracking calibration preceding the presentation of the stimuli (Fig. A1). Items were sorted by their number of infrequent words and characters, and short and long sentences, in ascending order. They were subsequently divided in four blocks and each block was shuffled randomly for each participant. Following this order, the experiment was carried out in two sessions of ten trials (two blocks, approximately one hour of reading). After the comprehension questions, a word association task was presented, where words were displayed (one by one) and the participant was required to write the first word that came to mind. For this task, five words were chosen randomly from the 150 most frequent words that are not propositions, verbs, articles (according to the corpus LexEsp (Sebastián Gallés et al., 1998)), and were not present in stories. The same five words were always presented for a given item. The goal of this task is to remove any lingering bias that may have remained from reading the story. The following trial did not begin until the participant agreed to it.

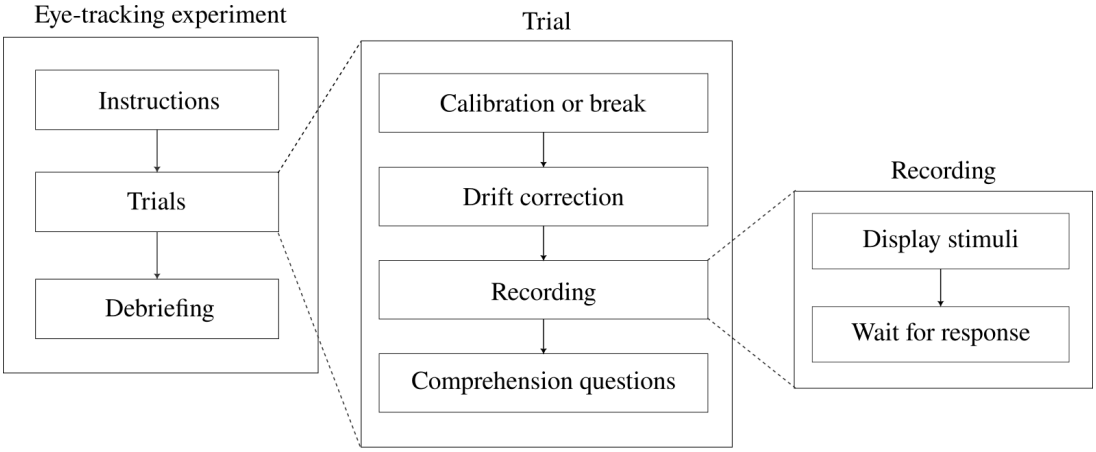

Figure A1: Depiction of the experiment setup, in which each trial consisted of a short story. The story was divided into screens and the participant was free to navigate forwards or backwards. At the end of each trial, the participant was required to answer comprehension questions about it.

A.3 Data processing

When processing a given trial, only fixations from the eye that had the least calibration error (as reported by the eye-tracker) were kept. Horizontal lines were drawn manually for each screen, indicating to what line of text each row of fixations belonged to. Very long and very short fixations (over 1000ms and under 50ms, respectively), as well as the first and last fixations in a screen, were automatically discarded. For a given word in a text line, its corresponding fixations are those whose x-coordinates fall within the word's surrounding blank spaces. Gaze measures were extracted from those words that were not the first or last words in a sentence or line and did not contain punctuation marks, dashes or numbers.

As the eye-tracker was used in binocular mode, fixations and saccades from both eyes were recorded. For data analysis, we only took into account the recordings from the eye that had the least calibration error as reported by the eye-tracker.

Each time a participant moved forward or backward to a different screen (pressed the right or left arrow), a message, with a timestamp, was logged by the eyetracker. These timestamps where utilized to divide the fixations by their corresponding screen. Some participants returned to a previous screen more than once, usually to get a better comprehension of the story so they could answer the questions accurately. In these cases, when the return was made by mistake (there are some fixations scattered across the screen in no particular order), the data is disposed of. However, when the return included re-reading some portion of the text, the data is kept, and the fixations are counted as regressions.

Once the data were curated and horizontal lines were drawn to decide to which text line each row of fixations corresponded to, fixation assignment to words followed. Given an item, this process is performed separately for each trial. As fixations are divided by screen, for each screen, text lines were first split into words by using blank spaces as separators. A subset of screen fixations is considered to belong to a given text line if their y-coordinate falls within the lower (included) and upper (excluded) bound of the corresponding horizontal lines.

In every screen, the first and last fixations are automatically discarded. Additionally, for each line, any regressive fixation between the first and the left-most is considered to be the result of oculomotor errors (i.e., *return sweep*) and is discarded. Fixations resulting from returning to the screen are numbered starting from the last fixation number on that screen. They are considered regressions if they fixate on the same words as the previous times. Fixations outside the scope of any word in the text are considered out of bounds.

B Non fine-tuned word pairs

When analyzing word pairs that were not present in the stimuli of the eye-tracking experiment (see 2.1), as expected, we found little to no change in the mean of the distributions (Fig. B2). In the case of Word2Vec, fine-tuning with Texts and its variation with gaze measures provided no difference whatsoever with respect to the baseline in all datasets. However, despite small changes to the

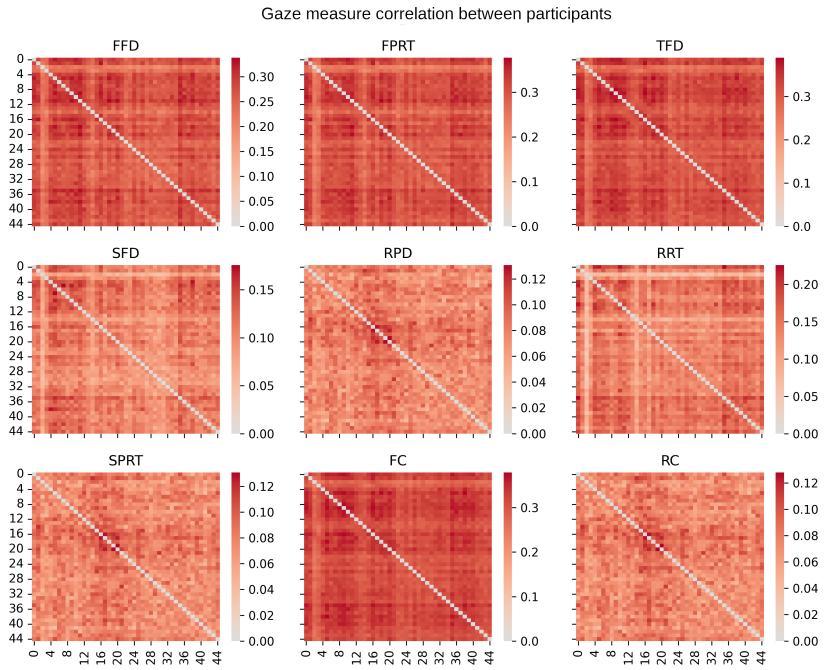

Figure A2: Correlation matrix for each gaze measure, averaged across items, between the 45 participants that read all items. Gaze measures were discretized in ten bins for each individual. FFD refers to First Fixation Duration; FPRT to First Pass Reading Time; TFD to Total Fixation Duration; SFD to Single Fixation Duration; RPD to Regression Path Duration; RRT to Re-Reading Time; SPRT to Second Pass Reading Time; FC to Fixation Count; RC to Regression Count.

mean, fine-tuning with Scanpaths provided significantly different values in all datasets compared to the baseline: in SimLex, the mean of the baseline was 0.4716 (s.e.m. 0.0086) and 0.4611 (s.e.m. 0.0087, $p < 0.0001$, ws. 275.0) for Scanpaths; in Abstract, 0.4389 (s.e.m. 0.0077) and 0.4190 (s.e.m. 0.0079, $p < 0.0001$, ws. 283.0); and, in Concrete, 0.4492 (s.e.m. 0.0083.0) and 0.4461 (s.e.m. 0.0083, $p < 0.0001$, ws. 1211.0).

With respect to AWD-LSTM, on the contrary, most variations provide significant differences in the distributions of the correlations with respect to the baseline in all datasets, with the exception of Scanpaths (with and without gaze measures) in Abstract ($p > 0.1$, ws. 2236.0 and 2423.5, respectively). When fine-tuning with Texts, with and without gaze measures, the mean of the distribution is slightly higher than the baseline: in SimLex, the mean of the baseline is 0.2991 (s.e.m. 0.0099) compared to 0.3076 and 0.3079 (s.e.m. 0.0098, $p < 0.0001$, ws. 442.0 and 413.0), respectively; in Abstract, 0.2594 (s.e.m. 0.0089) compared to 0.2728 (s.e.m. 0.0087, $p < 0.0001$, ws. 136.0 and 163.0); and, in Concrete, 0.2101 (s.e.m. 0.0098) compared to 0.2187 and 0.2192 (s.e.m. 0.0098, $p < 0.0001$, ws. 726.5 and 632.0). Scanpaths with and without gaze measures, on the other hand, slightly decreases the mean in SimLex (0.2877 and 0.2841, s.e.m. 0.01, $p < 0.005$, ws. 1067 and 1634, respectively) and in Concrete (0.1898 and 0.1952, s.e.m. 0.01, $p < 0.0001$, ws. 80.0 and 182.0, respectively).

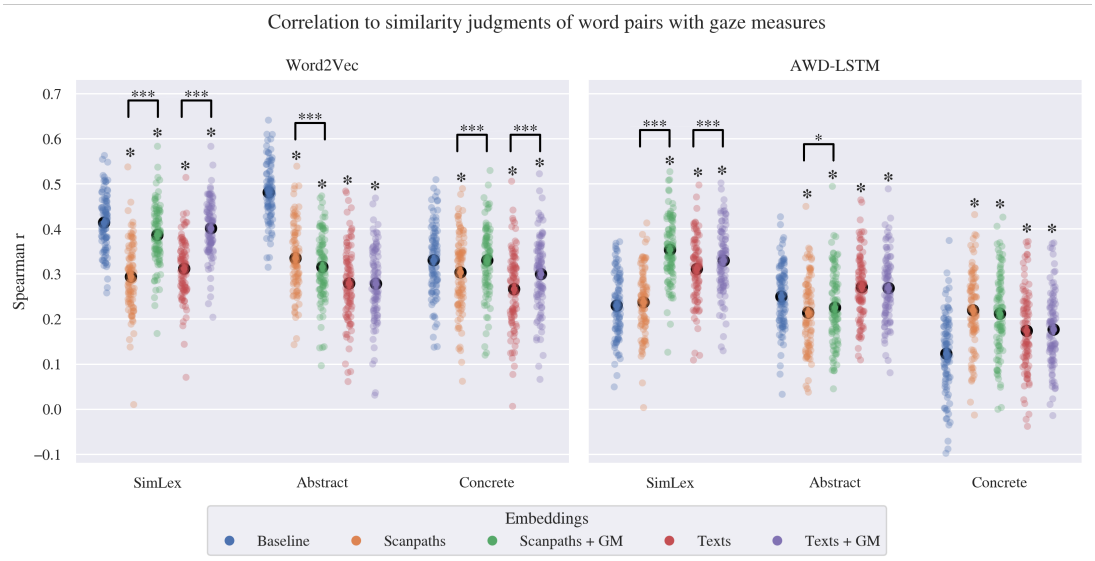

Figure B1: Distribution of the Spearman rank correlations of the cosine distances between word pairs that possess gaze measures with respect to human similarity judgments presented in Table 1.

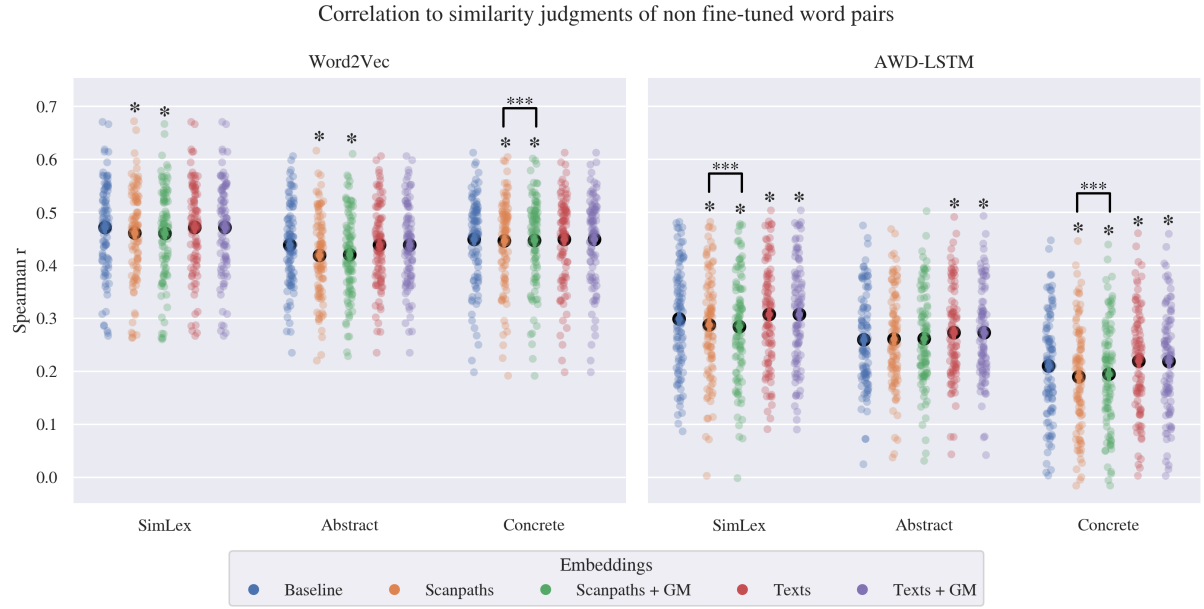

Figure B2: Spearman rank correlation of the cosine distance between word pairs that have not been fine-tuned (i.e., were not present in the stimuli of the eye-tracking experiment) with respect to human similarity judgments in three different datasets. 895, 1081, and 1431 word pairs were evaluated in total for SimLex, Abstract and Concrete, respectively, using random sampling with replacement of a hundred word pairs a hundred times. An asterisk above the strip plot indicates significance against the baseline. On the left are the results of the fine-tuning on Word2Vec and, on the right, on AWD-LSTM. Baseline refers to the models trained on the baseline corpus, whereas Scanpaths are the models fine-tuned on the text in the stimuli as read by the participants, and Texts are the models fine-tuned on the stimuli as is. GM refers to the addition of gaze measures to the training process.

Story	Author	Words	Fixations	Excluded fixations	Regressions	Skips
La noche de los feos	Mario Benedetti	544	25774	10290	8046	11234
Cómo funcionan los bolsillos	Valentín Muro	972	45815	11677	16176	19705
La máscara de la Muerte Roja	Edgar Allan Poe	572	26641	6805	9092	11974
Las fotografías	Silvina Ocampo	618	26686	8034	8580	12636
La salud de los enfermos	Julio Cortázar	667	34486	7596	12189	17953
Buenos Aires	Hernán Casciari	607	28813	6855	10368	12932
Wakefield	Nathaniel Hawthorne	693	31610	9034	10467	17397
Cómo funciona caminar en la nieve	Valentín Muro	1066	47302	10650	16245	20937
Ahora debería reírme, si no estuviera muerto	Angela Carter	606	25629	7124	7022	15558
El espejo	Haruki Murakami	628	29851	9597	9170	16788
Embarrar la magia	Facundo Alvarez Heduan	683	34749	12290	12143	14400
La lluvia de fuego	Leopoldo Lugones	640	30960	9236	10121	15979
Educar para escalar y bucear	Andrés Rieznik	599	27797	7472	9500	12621
El golpe de gracia	Ambrose Bierce	602	27629	7567	9540	14387
La gallina degollada	Horacio Quiroga	659	30188	8958	9825	15769
La canción que cantábamos todos los días	Luciano Lamberti	620	28299	7247	8418	15386
El almohadón de plumas	Horacio Quiroga	579	28063	9453	8301	15087
Una rosa para Emilia	William Faulkner	643	33946	8968	12007	16178
La de la Obsesión por la Patineta	Hernán Casciari	579	29200	8516	10044	13171
Total	-	**13218**	**623654**	**175481**	**206586**	**305949**

Table A1: List of stories employed in the eye-tracking experiment. Exclusion criteria for words (and their corresponding fixations) include being first or last in a sentence or screen line, or containing punctuation marks, dashes or numbers. Stories were divided in screens and participants were free to return to a previous screen. Fixations to words in a returning screen are counted as regressions.