Cuentos: Un corpus estandarizado de seguimiento ocular durante la lectura de textos narrativos para la investigación en Procesamiento del Lenguaje Natural y ciencias cognitivas

Fermin Travi (1,2), Bruno Bianchi (1,2), Diego Fernandez Slezak (1,2) y Juan E Kamienkowski (1,2,3)

(1) Laboratorio de Inteligencia Artificial Aplicada, Instituto de Ciencias de la Computación (Univ. de Buenos Aires – Consejo Nacional de Investigaciones Científicas y Técnicas), (C1428EGA), Buenos Aires, Argentina.

(2) Depto. de Computación (Fac. de Cs. Exactas y Naturales, Univ. de Buenos Aires), (C1428EGA), Buenos Aires, Argentina.

(3) Maestría de Explotación de Datos y Descubrimiento del Conocimiento (Univ. de Buenos Aires), (C1428EGA), Buenos Aires, Argentina.

El seguimiento ocular es un método bien establecido para estudiar los procesos de lectura. En estos, nuestra mirada salta de palabra en palabra, muestreando información de manera casi secuencial. El tiempo dedicado a cada palabra, junto con los patrones de omisión o revisión, proporciona indicadores de los procesos cognitivos durante la comprensión del texto. Sin embargo, pocos estudios se han centrado en el español, donde los datos empíricos siguen siendo escasos y se sabe poco sobre cómo los hallazgos de otros idiomas se traducen al comportamiento de lectura en español. En paralelo, estudios recientes han comenzado a demostrar que la integración de datos oculomotores resulta una vía prometedora para alinear de forma más precisa los modelos computacionales de lenguaje con el procesamiento cognitivo humano y, en última instancia, mejorar el desempeño de dichos modelos. A pesar de que la mayoría de estos estudios han sido realizados en idioma Inglés, los experimentos realizados con datos de movimientos oculares en arquitecturas clásicas de Procesamiento de Lenguaje Natural (PNL) con datos en español muestran un camino promisorio [1], marcando la necesidad de contar con más cantidad y variedad de datos para el entrenamiento de los modelos.

En este trabajo, publicado en la revista Scientific Data, especializada en la descripción y presentación de corpus científicos, presentamos el mayor corpus de seguimiento ocular en español disponible públicamente hasta la fecha [2], que comprende lecturas de historias autoconclusivas de 113 hablantes nativos (edad media 23,8; 61 mujeres, 52 hombres). El corpus comprende tanto historias largas (3300 ± 747 palabras, 11 lecturas por ítem en promedio) como historias cortas (795 ± 135 palabras, 50 lecturas por ítem en promedio), proporcionando una amplia cobertura de escenarios de lectura natural con más de 940.000 fijaciones que cubren cerca de 40.000 palabras (8.500 palabras únicas).

El corpus se encuentra disponible a través del enlace https://doi.org/10.6084/m9.figshare.28311908. Asimismo, las herramientas y el paquete de código necesarios para la validación y extracción automática de las medidas de seguimiento ocular durante la lectura se encuentran disponibles en https://github.com/NeuroLIAA/reading-et. Estos recursos permiten también expandirlo fácilmente, agregando nuevos sub-corpus, nuevas métricas, o extendiendo las métricas existentes (como la predictibilidad a todo el corpus).

Este recurso integral ofrece oportunidades para investigar los patrones de movimiento ocular en español, explorar procesos cognitivos específicos del idioma, examinar fenómenos lingüísticos del español y desarrollar algoritmos computacionales para la investigación de la lectura y aplicaciones de procesamiento del lenguaje natural. La integración de señales cognitivas representa una tendencia de investigación clave para enriquecer los modelos de lenguaje, ya que permite optimizar su entrenamiento, mitigar posibles alucinaciones y mejorar de manera significativa su alineación con el comportamiento y las intenciones de los usuarios [3].

[1] Travi, F., Leclercq, G. A., Slezak, D. F., Bianchi, B., & Kamienkowski, J. E. (2025, May). *Exploring the Integration of Eye Movement Data on Word Embeddings*. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (pp. 55-65).

[2] Travi, F., Bianchi, B., Slezak, D. F., & Kamienkowski, J. E. (2026). *Cuentos: A Large-Scale Eye-Tracking Reading Corpus on Spanish Narrative Texts*. Scientific Data. **13**, 434

[3] López-Cardona, Á., Segura, C., Karatzoglou, A., Abadal, S., & Arapakis, I. (2025, May). *Seeing eye to AI: human alignment via gaze-based response rewards for large language models*. In International Conference on Learning Representations (Vol. 2025, pp. 26111-26135).

https://doi.org/10.1038/s41597-026-06798-z

Scientific Data; Journal Impact Factor: 6.9 (2024); 5-year Journal Impact Factor: 8.7 (2024); Q1