**Métodos distribucionales y visualización aplicados a la semántica léxica**

Mariana Montes

Universidad de Lovaina (KU Leuven, Bélgica)

[mariana.montes@kuleuven.be](mailto:mariana.montes@kuleuven.be)

Una de las fuentes de lenguaje auténtico que se prestan a estudios empíricos del lenguaje es un corpus. La disponibilidad de grandes cantidades de texto espontáneo en formato digital abre las puertas para análisis a gran escala, lo que implica mayor riqueza de información pero también la necesidad de técnicas estadísticas avanzadas para procesar esta información. A su vez, existe una brecha entre los aspectos observacionales que una computadora puede extraer y manipular y los aspectos conceptuales: la interpretación que le atribuimos a los fenómenos lingüísticos. Esta brecha es particularmente clara en el caso de la semántica léxica: la búsqueda y descripción del *significado* de ítems léxicos. En tanto una máquina no puede *entender* el lenguaje de la misma manera que los humanos, ¿cómo podemos aprovechar sus habilidades computacionales y de identificación de patrones para agilizar nuestros análisis semánticos?

En el marco de un [proyecto](https://www.arts.kuleuven.be/ling/qlvl/projects/current/nephological-semantics) dedicado al desarrollo de herramientas para estudios semánticos con métodos distribucionales (Heylen et al., 2015, 2012; Lenci, 2018), esta presentación es la síntesis de una tesis doctoral (Montes, 2021) que examina la aplicabilidad de espacios vectoriales al estudio de la estructura polisémica de ítems léxicos.

Concretamente, extraemos información distribucional de un corpus (frecuencias) para representar las ocurrencias de distintos ítems léxicos, de forma tal que ocurrencias similares (idealmente, instancias de sentido equivalente) son proyectadas como puntos cercanos en un gráfico. Por ejemplo, el verbo neerlandés *huldigen* puede significar ‘rendir homenaje’ o ‘sostener (una opinión)’ dependiendo del objeto directo. La pregunta es: ¿podemos extraer información de un corpus de forma tal que, automáticamente, obtengamos grupos separados de ocurrencias que corresponden a los distintos sentidos? De no ser así, ¿qué información sí puede ser extraída de un corpus, y cómo nos puede ayudar en la descripción semántica de ítems léxicos?

En esta charla presentaré (1) la técnica de modelos vectoriales de nivel de caso (Heylen et al., 2015; Schütze, 1998), (2) la visualización interactiva desarrollada dentro del equipo de investigación para explorar los resultados (Montes & Heylen, 2022; Montes & Wielfaert, 2021) y (3) las conclusiones más importantes.

A través de un estudio exhaustivo de 32 sustantivos, verbos y adjetivos en neerlandés, mostraré cómo estos métodos logran capturar patrones colocacionales, pero que la medida en que éstos caracterizan un sentido en términos definicionales depende del comportamiento distribucional específico de cada palabra.

Bibliografía

Heylen, K., Speelman, D., & Geeraerts, D. (2012). Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. *Proceedings of the Eacl 2012 Joint Workshop of LINGVIS & UNCLH*, 16–24.

Heylen, K., Wielfaert, T., Speelman, D., & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, *157*, 153–172.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*(1), 151–171.

Montes, M. (2021). *Cloudspotting: Visual analytics for distributional semantics* [PhD Dissertation]. KU Leuven.

Montes, M., & Heylen, K. (2022). Visualizing Distributional Semantics. En D. Tay & M. X. Pan (Eds.), *Data Analytics in Cognitive Linguistics. Methods and Insights*. Mouton De Gruyter.

Montes, M., & Wielfaert, T. (2021). *QLVL/NephoVis: Altostratus* (v1.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.5116843

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, *24*(1), 97–123.