La identificación automática de significados gramaticales en datos de corpus mediante Latent Class Analysis

Malte Rosemeyer
Freie Universität Berlin

La variación en el formato morfosintáctico de los enunciados frecuentemente puede explicarse en términos de diferencias de significado (Bybee, 2010: 165). Por ejemplo, en español, las perífrasis *tener que* + infinitivo, *deber* + infinitivo y *deber de* + infinitivo pueden expresar significados modales deónticos (1) o epistémicos (2). Se supone que el uso de *tener que* + infinitivo es más probable con lecturas deónticas que el uso de *deber* + infinitivo y especialmente *deber de* + infinitivo (Blas Arroyo, 2011; Rosemeyer, 2017). En cambio, el uso de las perífrasis con el auxiliar *deber* sería más probable con lecturas epistémicas.

(1) a. *Ten-go que cant-ar.*
b. *Deb-o cant-ar.*
c. *Deb-o de cant-ar.*

(2) a. *Tien-e que ser Juan.*
b. *Deb-e ser Juan.*
c. *Deb-e de ser Juan.*

Identificar los significados de elementos gramaticales en contexto es un desafío importante para los estudios de variación gramatical en los datos de corpus. Este estudio propone una solución novedosa a este problema. Describo los significados situados de elementos gramaticales como construcciones latentes (*latent constructs*). Los constructos latentes son variables no observables pero medibles en términos de indicadores que representan el constructo subyacente (Nylund-Gibson and Choi, 2018). Por tanto, los significados situados no pueden observarse directamente, sino que deben inferirse de la forma en que se comportan los hablantes. Estos indicadores son características del contexto lingüístico y no lingüístico.

Utilizo el Análisis de Clases Latentes (Latent Class Analysis, LCA) para establecer una tipología de significados gramaticales basada en datos para las tres perífrasis modales ilustradas en (1)-(2) para mostrar cómo se puede utilizar el LCA para identificar significados gramaticales no observados en función de su distribución respecto de un conjunto de predictores contextuales. Luego comparo esta tipología con la clasificación manual de los datos en términos de modalidad. Para realizar este análisis, utilizo datos de entrevistas sociolingüísticas habladas de España (Preseea, 2014).

Mis resultados muestran que (a) los significados situados identificados por el LCA no se corresponden directamente con los significados modales que comúnmente se supone que rigen la variación entre las tres perífrasis, y (b) la tipología de significados basada en datos es mejor para explicar la variación entre estas perífrasis. Mi análisis también considera la relevancia del estatus socioeconómico para esta variación y muestra que la probabilidad de realización de ciertos tipos de significados situados por parte de los hablantes está regida por su estatus socioeconómico.

Bibliografía

Blas Arroyo, José Luis (2011). *Deber* (*de*) + infinitivo: ¿un caso de variación libre en español? Factores determinantes en un fenómeno de alternancia sintáctica". *Revista de Filología Española* 91(1): 9-42.

Bybee, Joan L. (2010). *Language, Usage, and Cognition*. Cambridge, New York, Cambridge University Press.

Nylund-Gibson, Karen and Andrew Young Choi (2018). Ten frequently asked questions about Latent Class Analysis. *Translational Issues in Psychological Science* 4: 440-461. 10.1037/tps0000176

Preseea (2014). *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares, Universidad de Alcalá. Available online at http://preseea.linguas.net. Last access 6 January 2020.

Rosemeyer, Malte (2017). La historia de las perífrasis *deber* / *deber de* + INF: variación, norma y géneros textuales. *La gramática en la diacronía. La evolución de las perífrasis verbales modales en español*. Mar Garachana. Madrid, Frankfurt a.M., Iberoamericana, Vervuert**:** 147-195.