Putting semantics back on the map: enriching grammatical alternation research with distributional semantics

Chiara Paolini, Hubert Cuyckens, Stefania Marzo, Dirk Speelman, & Benedikt Szmrecsanyi
KU Leuven
chiara.paolini@kuleuven.be, hubert.cuyckens@kuleuven.be, stefania.marzo@kuleuven.be,
dirk.speelman@kuleuven.be, benedikt.szmrecsanyi@kuleuven.be

Keywords: variation, synonymy, isomorphism, distributional semantics

This study focusses on the ditransitive (1a) and prepositional (1b) dative variants as semantically/functionally broadly interchangeable syntactic variants in English (see Bresnan et al. 2007).

(1) a. [The child]$_{subject}$ [gave]$_{verb}$ [her mother]$_{recipient}$ [a flower]$_{theme}$
b. [The child]$_{subject}$ [gave]$_{verb}$ [a flower]$_{theme}$ [to her mother]$_{recipient}$

This alternation is extremely well researched. However, insufficient attention has been paid to how lexical-semantic properties of the embedding linguistic context influence linguistic choice-making. In most cases, the reason is rather practical: manually annotating for top-down semantic properties (such as constituent animacy, typically analysed as a predictor in dative alternation research) is labor-intensive and time-consuming. Besides, annotation for these top-down properties is limited in the extent to which it can represent lexical-semantic richness and variation. Thus, the bulk of the literature on grammatical alternations relies on traditional, top-down formal predictors such as pronominality, or constituent length. Our aim is thus to determine the importance of semantic properties of the lexical context for predicting variant choice. Whereas – in line with the variationist methodology – we assume broad functional equivalence between the dative variants themselves, we are interested in the extent to which semantics plays an indirect role via the lexical material in the constituent slots. Ultimately, therefore, we address the extent to which two foundational principles in Cognitive Linguistics, the Principle of Isomorphism (Haiman 1980) and the Principle of No Synonymy (Goldberg 1995), can be reconciled with the rich variationist literature on the existence, ubiquity, and systematicity of grammatical variation between alternate ways of saying the same thing (Labov 1972).

On the technical plane, we experiment with a fully automatic, bottom-up method to model constituent semantics, which involves creating semantic predictors using distributional models of meaning (Lenci 2018). We specifically assess the semantics of the heads of the noun phrases taking the role of theme and recipient (e.g., flower and mother in (1)) via type-level semantic vector space modelling. The models were trained on the spoken COCA (Davies 2019, ~127 million words). Based on the resulting distance matrices, we automatically clustered theme and recipient heads separately, and obtained groupings of semantically-related types. Recipient heads clustered into rather coherent groupings related to family roles, job titles, economics, and law terminology, as well as anaphoric pronouns. Conversely, theme heads yield a wider range of semantic groupings, including words related to the labour market and household items. We then used these clusters as categorical predictors in mixed-effects binary logistic regression and conditional random forest models.

The statistical analysis is based on N = 1,170 dative observations with the dative verb give drawn from the dataset used in Bresnan et al. (2007), which in turn is derived from the Switchboard corpus of US American English (Godfrey and Holliman 1993).

In an effort to combine the top-down grammatical-oriented tradition of analysis with bottom-up lexical-semantic data, we then fitted a number of regression models with traditional formal predictors and vector-space-modelling-derived predictors. Analysis suggests that while bottom-up semantic clusters have significant predictive power, they are outperformed by traditional predictors, such as constituent weight.

References

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. “Predicting the Dative Alternation.” *Cognitive Foundations of Interpretation*, 69–94.

Davies, Mark. 2019. “The Corpus of Contemporary American English (COCA): One Billion Words, 1990-2019.” 2019. https://www.english-corpora.org/coca/.

Godfrey, John J., and Holliman, Edward. 1993. “Switchboard-1 Release 2.” Linguistic Data Consortium. https://doi.org/10.35111/SW3H-RW02.
Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
Haiman, John. 1980. “The Iconicity of Grammar: Isomorphism and Motivation.” *Language* 56 (3): 515. https://doi.org/10.2307/414448.
Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Philadelphia Press.
Lenci, Alessandro. 2018. “Distributional Models of Word Meaning.” *Annual Review of Linguistics* 4 (1): 151–71. https://doi.org/10.1146/annurev-linguistics-030514-125254.