Centre Language – Variation – Multilingualism

Linguistic Lecture Series Summer Term 2024

Malte Rosemeyer:

Data-driven identification of situated meanings in corpus data using Latent Class Analysis

Abstract:

Variation in the morphosyntactic format of utterances can frequently be explained in terms of meaning differences (Bybee, 2010: 165). For instance, in Spanish, the periphrases *tener que* + infinitive 'have to', *deber* 'must' + infinitive and *deber de* 'must' + infinitive can express deontic (1) or epistemic modal meanings (2). *Tener que* + infinitive is assumed to be more likely to be used with deontic readings than the *deber* + infinitive and especially *deber de* + infinitive. The reverse is true for epistemic readings.

(1) a. *Ten-go* *que* *cant-ar.*
have-PRS.IND.1SG that sing-INF
b. *Deb-o* *cant-ar.*
must-PRS.IND.1SG sing-INF
c. *Deb-o* *de* *cant-ar.*
must-PRS.IND.1SG of sing-INF
'I have to sing.'

(2) a. *Tien-e* *que* *ser* *Juan.*
have-PRS.IND.3SG that be.INF Juan
b. *Deb-e* *ser* *Juan.*
must-PRS.IND.3SG be.INF Juan
c. *Deb-e* *de* *ser* *Juan.*
must-PRS.IND.3SG of be.INF Juan
'That must be Juan.'

Identifying the meanings of grammatical elements in context is a major challenge for corpus-linguistic studies of grammatical variation. This study proposes a novel solution to this problem. I describe the situated meanings of grammatical elements as as latent constructs.

SVM ZENTRUM Sprache | Variation Mehrsprachigkeit

Universität Potsdam

Latent constructs are variables that non-observable but measurable in terms of indicators that represent the underlying construct (Nylund-Gibson and Choi, 2018). Thus, situated meanings cannot be observed directly but need to be inferred from the way that speakers behave. These indicators are features of the linguistic and non-linguistic context.

I use Latent Class Analysis (LCA) to establish a data-driven typology of grammatical meanings for the three modal periphrases illustrated in (1)-(2) to show how LCA can be used to identify unobserved grammatical meanings based on their distribution in terms of a set of contextual predictors. I then compare this typology to manual classification of the data in terms of modality. In conducting this analysis, I use data from spoken sociolinguistic interviews (Preseea, 2014).

My findings show that (a) the situated meanings identified by the Latent Class Analysis do not directly correspond to the modal meanings that are commonly assumed to govern the variation between the three periphrases, and (b) the data-driven typology of meanings is better in explaining the variation between these periphrases. My analysis also considers the relevance of socioeconomic status for this variation and shows that certain types of situated meanings are more likely to expressed by speakers with a higher socioeconomic status.

References

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.