# Comparing memory-based learning and regression approaches in the explanation of syntactic variation and change in Belgian and Netherlandic Dutch.

Stefan Grondelaers (Radboud University Nijmegen), Antal van den Bosch (Radboud University Nijmegen), Dirk Speelman (University of Leuven), Roeland van Hout (Radboud University Nijmegen)

In this paper, we investigate the merits of memory-based learning as a supplement to the regression techniques which are current in socio-syntactic analysis. Memory-based learning (MBL, Daelemans & Van den Bosch 2005) is an algorithm which predicts constructional choice on the basis of similarity with a training set of stored examples.

Our case study focuses on the post-verbal distribution of existential *er* “there” in Dutch locative inversion constructions (such as *In de asbak ligt (er) een sigarenpeuk* “In the ashtray (there) lies a cigar butt”). *Er*-insertion is sensitive to many internal and external variables, including national variation between Belgian and Netherlandic Dutch.

Previous attempts to account for the differences between the Belgian and Netherlandic distribution of *er* (Grondelaers et al. 2008) relied on regression analyses which built on higher-order predictors pertaining to the syntactic, semantic and discourse properties of the adjunct and the verb. While these analyses demonstrated that Netherlandic preferences were easier to model than Belgian preferences, they did not reveal to what extent the identity and combinability of the raw lexemes in the constructions conditioned the presence of *er*.

MBL offers precisely this advantage. Theijssen (2013) found that an MBL-model rivalled a regression analysis of the dative alternation (*I gave the ball to her* vs. *I gave her the ball*) on account of the strong preference of specific verbs for one of both options. In order to find out whether lexical feature representations also suffice to account for more complex syntactic variables such as *er*-insertion, we carried out a series of MBL-experiments in which we trained the Belgian and Netherlandic classifier on a small dataset of manually annotated examples (n < 1.000) and on a large dataset of examples drawn from automatically parsed corpora (n > 100.000).

While *er*-insertion in Netherlandic Dutch turned out to be easy to learn (even on the basis of the small dataset), learning to insert *er* in Belgian Dutch proved much harder, and was only moderately successful on the basis of the large set. Crucially, the Netherlandic MBL model rivalled the performance of the previous regression analysis. In Belgian Dutch, by contrast, the marked superiority of the regression suggests a selection process relying on abstract features which are difficult to learn from lexical input.

Diachronically, the reported data reveal that the more advanced linguistic standardization of Netherlandic Dutch (Grondelaers & Van Hout 2011) also manifests itself in a spontaneous reduction of the number of possible collocations between *er* and specific lexemes in the adjunct and verb slots. A synchronic conclusion seems to be that *er*-insertion in Netherlandic Dutch has become exemplar-based. Methodologically, our findings challenge the undisputed supremacy in sociolinguistics of regression techniques which underspecify the impact of uncategorized lexical information. We advocate the combination of both regression- and memory-based learning tools to access *all* the factors which determine constructional choice.