What is being modelled?

In this talk I would like to take a look at recent data-driven methods from a more distant perspective. Rather than presenting a single case in detail, I will take pause and look at a number of approaches to a number of case studies, with the specific question in mind: what is it exactly that we are modelling here? Specifically, I would like to talk in a bit more detail about the relationship between the macro-level and lower levels of granularity by looking at a number of case studies of syntactic change.

The first case I would like to discuss relates to the recent attempts at quantifying grammaticalization beyond merely counting frequencies. In Petré & Van de Velde (2018) a method was proposed for measuring grammaticalization at the level of the individual data point, by combining several semantic and formal features into an overall grammaticalization score for each data point, allowing more fine-grained diachronic study of grammaticalization over time. Some other recent studies, such as Saavedra 2019 or Dekalo 2021 have proposed related measures, but only at the aggregate level. I present a follow-up study of Petré & Van de Velde looking at the same case study (*be going to* INF), but using slightly different criteria. While almost identical results materialize at the aggregate level of the community, the analysis of individual lifespan change greatly differs from the original study. Specifically, where the original study suggested that generations living in the middle of a change show more lifespan change, this is less clear in the replication study. Additionally, there was little overlap in which authors came out as significant changers between the two studies. This raises interesting questions about the different status of the aggregate results from the individual results. One interpretation, I will argue, is that frequency is still overshadowing the qualitative variables in this method.

A second case I will discuss is the research by Budts (2020), who uses artificial neural networks to trace the development of *do*, and measure its similarity to modal auxiliaries. Distributional methods like these are typically implemented at the aggregate level, for the obvious reason that individual-level data are still generally too scarce for such methods. What exactly does it imply that the model is able to predict language choices on the basis of aggregate data? And are the areas in which the highest degree of uncertainty is shown explicable as indicative of widely shared polysemy, or rather indicative of split usage dependent on individual usage, genre, or community of practice? What would we need to change about these models to make them more fine-grained in this respect?

While this may sound a bit pessimistic, the above models also have clear benefits, for instance if compared to certain types of regression modelling. Mixed models regression only has limited tools to deal with the non-randomness of the individual level. More generally, regression models struggle to see variables as reinforcing each other rather than competing with each other. However, because distributional methods do not easily allow you to see what exactly they have learned in terms of variables, it is not straightforward to compare these outcomes.