Markov models for diachronic corpus studies
(Freek Van de Velde & Isabeau De Smet)

In corpus linguistics, quantitative modeling of diachronic trends is often achieved by generalized linear models, most often with the logit link (a.k.a. logistic regression), and time as a predictor and competing variants as the outcome variable. Advantages of this approach are: (1) the trajectory follows a sigmoid curve, which is known to underly many changes (Denison 2003; Blythe & Croft 2012, among many others), (2) the model allows for multi-variate control, with main effects and interactions, (3) the model can accommodate numerical and categorical predictors, (4) the model can accommodate extra-linguistic and intra-linguistic factors.

These advantages should not distract us from the fact that adding time as a predictor in such models is not always straightforward (Hosmer et al. 2008: 2; Koplenig & Müller-Spitzer 2016; Koplenig 2017; Van de Velde & Petré 2020). One of the problems is that the transition between competing variants may show a wavering back-and-forth movement, instead of a gradual transition. Take, for instance, the Germanic synthetic preterite, which comes in two inflectional variants: the strong inflection, relying on Ablaut, e.g. *drive-drove*, and the weak inflection, relying on a dental suffix, e.g. *play-played*. While there is a long-term trend of 'weakening' of the strong Germanic preterite (Lieberman et al. 2007; Carroll et al. 2012; De Smet & Van de Velde 2019), there are also verbs that display a trend in the opposite direction (Cuskley et al. 2014; De Smet 2021), and the smooth sigmoid curve is hard to discern in the aggregate data. For these less tidy kinds of changes, we propose an alternative approach, in the form of a technique known as a Markov model in continuous time (Krylov 1995).

We looked at the morphological shifts in Germanic preterites, focusing on Dutch. We collected a database of 14314 Dutch preterites from 800AD to 2000AD, of verbs that in the course of time have been attested in the strong form. Using a so-called 'competing risks illness-death model', we allow transitions from strong to weak and vice versa, and also integrate the possibility of lexical death. Our results (see Van de Velde & De Smet 2021) show that the model brings out the known effects of frequency and ablaut-pattern, and can be used for fine-grained predictions.

Carroll, R., R. Svare & J. Salmons. 2012. 'Quantifying the evolutionary dynamics of German verbs'. *Journal of Historical Linguistics* 2: 153-172.

Cuskley, C., M. Pugliese, C. Castellano, F. Colaiori, V. Loreto & F. Tria. 2014. 'Internal and external dynamics in language: evidence from verb regularity in a historical corpus of English'. *PLoS ONE* 9(8). e102882.

Denison, D. 2003. 'Log(ist)ic and simplistic S-curves'. In: R. Hickey (red.), *Motives for language change*. Cambridge: Cambridge University Press, 54-70.

De Smet, I. 2021. *De sterke werkwoorden in het Nederlands. Een diachroon, kwantitatief onderzoek*. PhD Dissertation, KU Leuven.

De Smet, I. & F. Van de Velde. 2019. 'Reassessing the evolution of West Germanic preterite inflection'. *Diachronica* 36(2): 139-179.

Hosmer, D.W., S. Lemeshow & S. May. 2008. *Applied survival analysis: regression modeling of time-to-event data*. 2nd edn. New York: John Wiley.

Koplenig, A. 2017. 'Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions'. *Digital Scholarship in the Humanities* 32(1): 159-168.

Koplenig, A. & C. Müller-Spitzer. 2016. 'Population size predicts lexical diversity, but so does the mean sea level – why it is important to correctly account for the structure of temporal data'. *PLoS ONE* 11(3): e0150771.

Krylov, Y.K. 1995. 'A Markov model for the evolution of lexical ambiguity'. *Journal of Quantitative Linguistics* 2(1): 19-26.

Lieberman, E., J.-B. Michel, J. Jackson, T. Tang & M.A. Nowak. 2007. 'Quantifying the evolutionary dynamics of language'. *Nature* 449: 713-716.

Van de Velde, F. & P. Petré. 2020. 'Historical linguistics'. In: D. Knight & S. Adolphs (eds.), *The Routledge handbook of English language and digital humanities*. London: Routledge. 328-359.

Van de Velde, Freek & Isabeau De Smet. 2021. 'Markov models for multi-state language change'. *Journal of Quantitative Linguistics* (published ahead of print).