An alternative to generalized linear models: Markov models for multi-state language change

Freek Van de Velde

Isabeau De Smet

RU Quantitative Lexicology and Variational Linguistics

Generalized linear mixed models

- Main workhorse in quantitative (diachronic) corpus linguistic
- Has: multiple explanatory vars, link function, linear part, random intercepts, random slopes ...
- E.g. incoming mutant replacing old mutant, with a β parameter for Time var.:

$$\ln\Big(odds\big(x_j = Mutant\big)\Big) = \beta_0 + \sum_{i=1}^{n} \beta_i v_{i,j} + u_{0,j} + u_{i,j}\, v_{i,j}$$

$$logit\big(x_j = Mutant\big) = \beta_0 + \sum_{i=1}^{n} \beta_i v_{i,j} + u_{0,j} + u_{i,j}\, v_{i,j}$$

Problems

- Time is not your simple stochastic variable
- Non-independence: autocorrelation
- Generalized linear models assume a smooth transition, not a wavering back-and-forth process
- Sampling is often unequally distributed over the observed time span
- Mutant hosts (often treated as random factors) may fall out of use by lexical replacement, or may display a skewed distribution

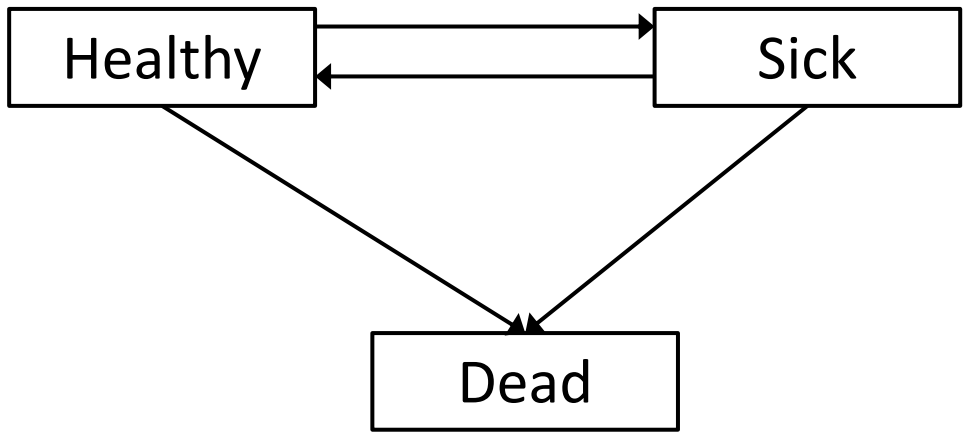
Solutions

- Non-independence: autocorrelation
  - Add control variable for previous value (Baayen & Milin 2010; De Smet & Van de Velde 2020)
  - Use advanced time-series modelling (Koplenig 2017; Koplenig et al. 2016; Van de Velde & Petré 2020)
- Generalized linear models assume a smooth transition, not a wavering back-and-forth process
  - Smoothen anyway (standard approach)
  - Use growth-curve analysis or additive models (GAMMs) (Winter & Wieling 2016)
- Sampling is often unequally distributed over the observed time span
  - Estimate from more densely populated parts of the model (standard approach)
  - Bin in coarser grained time variables (standard approach)
- Mutant hosts (often treated as random factors) may fall out of use by lexical replacement, or may display a skewed distribution
  - Assume robustness of random effects in non-normal distributions (standard approach)
  - Bin underpopulated random factors (standard approach)
  - Use regularization techniques (Van de Velde & Pijpops, forthc.)

Other solution

- Multi-state Markov Model
- A case of Survival Analysis
  - Takes the special nature of the time var. seriously
  - Deals with lexical replacement
  - Does not assume a smooth, (generalized) linear trend
  - Does not assume continuous sampling (as opposed to many time-series analyses)

Multi-state Markov Model



Case-study: Germanic preterites (Dutch)

- Germanic languages have two morphological strategies for building preterites (not counting analytic perfects, *he has written a book*):

1. Strong inflection
   - English *sing – sang*
   - Ablaut, based on Indo-European aspectual system (perfect > preterite)

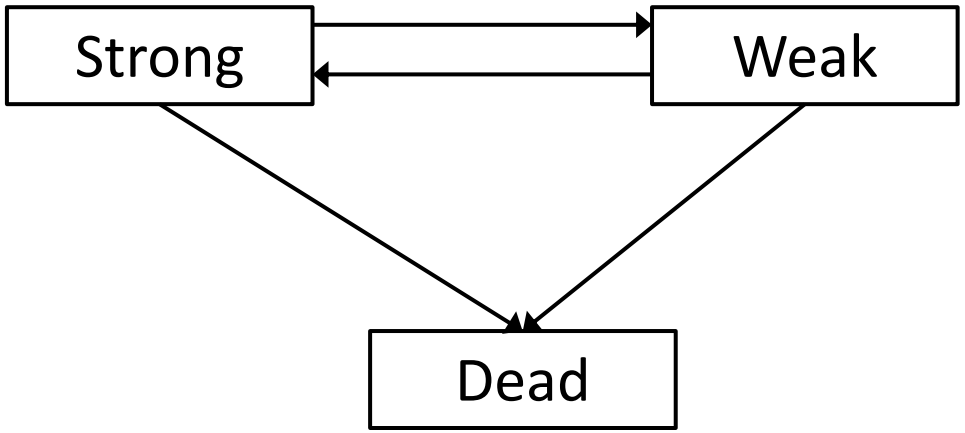
2. Weak inflection
   - English *work – worked*
   - Dental suffix, based on a analytic formation [VERB + $*d^{h}eh_{1}$-, $*d^{h}oh_{1}$- ('did')]

Case-study: Germanic preterites (Dutch)

- Diachrony has been studied intensely:
  - E.g. Anderwald, 2012; Cuskley et al., 2014 on 19th-century English; Lieberman et al., 2007 on Old English to Present-day English, Carroll et al., 2012 on Old High German to Present-day German, De Vriendt, 1965 on 16th-century Dutch; De Smet & Van de Velde, 2019, 2020, De Smet 2021 on 9th-century to 20th-century Dutch.
- Long-term drift, over many centuries
- Strong to weak, weak to strong, lexical death

Multi-state Markov Model

- Each verb (type) is a 'patient'



Data and methods

- Dutch preterites
- 285 verb types, 14314 tokens
- 800AD to 2000AD
- Based on database De Smet (2021)
- Covariates that are known to play a role:
  - Frequency (based on $\log_{10}$ token frequency of preterite of verb stem), ternary: high, medium, low
  - Ablaut pattern (pres – pret – part), ternary: ABB, ABA, ABC
- See Van de Velde & De Smet (2021) for mathematics and further data curation
- Theoretical application to linguistics (Krylov 1995)
- R package msm (Jackson, 2011, 2019)

		to:		
from:		strong	weak	dead
	strong	10741	265	38
	weak	240	3009	17
	dead	0	0	0

Results

covariate		transition	coefficient	confidence interval
	low	(reference level)		
FREQUENCY	mid	from strong to weak	1.06	[0.74, 1.54]
		from strong to dead	0.03	[0.01, 0.10]
		from weak to strong	2.38	[1.56, 3.64]
		from weak to dead	0.20	[0.06, 0.66]
	high	from strong to weak	0.37	[0.24, 0.57]
		from strong to dead	<0.01	[<0.01, >100.00]
		from weak to strong	2.52	[1.56, 4.08]
		from weak to dead	<0.01	[0.00, ∞]
VOWEL PATTERN	ABB	(reference level)		
	ABC	from strong to weak	1.44	[1.00, 2.08]
		from strong to dead	1.31	[0.58, 2.93]
		from weak to strong	0.46	[0.32, 0.67]
		from weak to dead	0.34	[0.06, 1.99]
	ABA	from strong to weak	3.80	[2.86, 5.06]
		from strong to dead	0.93	[0.36, 2.41]
		from weak to strong	0.66	[0.49, 0.89]
		from weak to dead	0.76	[0.25, 2.32]

Transition probability matrix for $t = 100$ (years).

		to:		
from:		strong	weak	dead
	strong	0.861	0.139	<0.001
	weak	0.510	0.490	<0.001
	dead	0.000	0.000	1.000

Transition probability matrix for $t = 500$ (years).

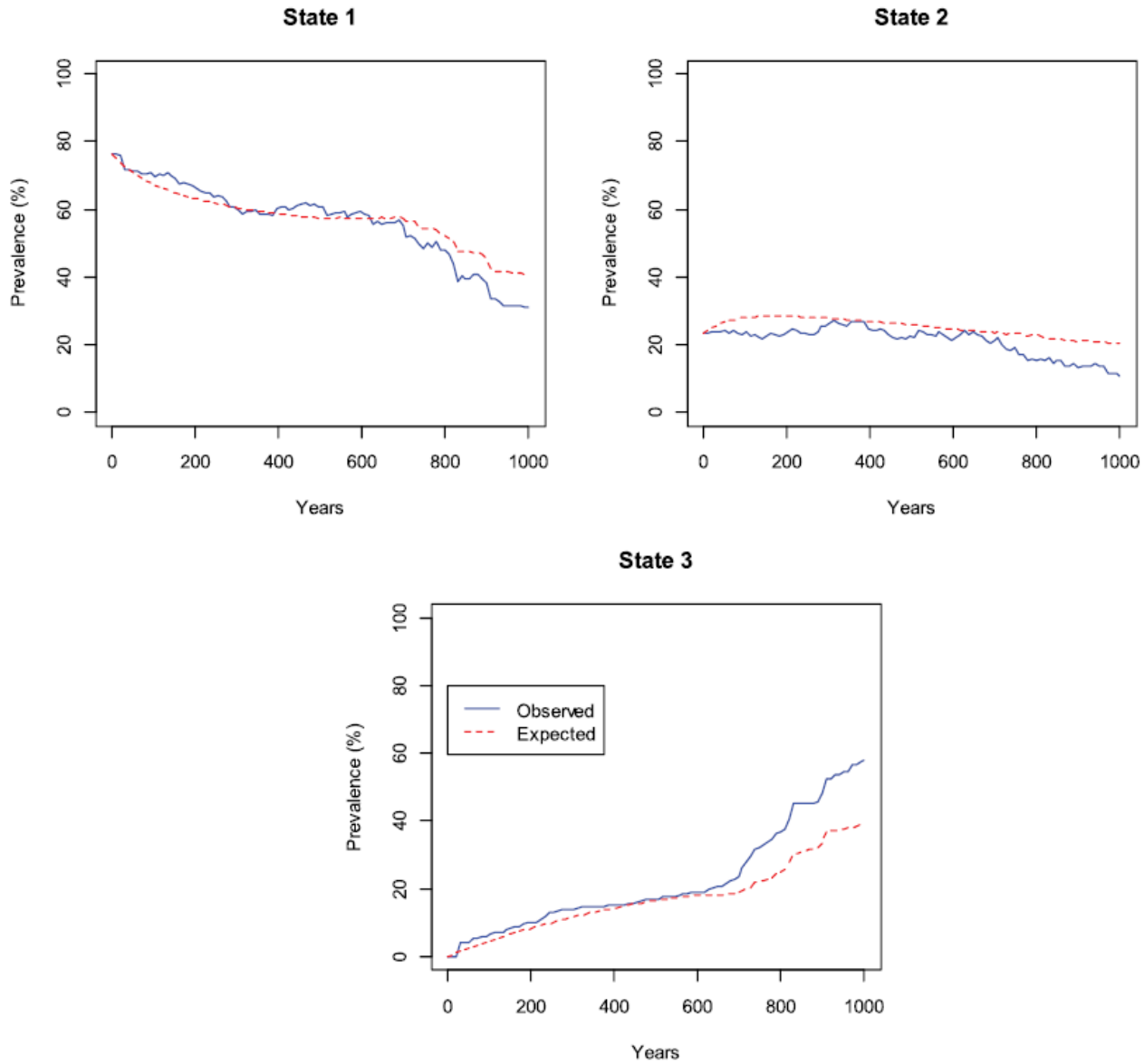
		to:		
from:		strong	weak	dead
	strong	0.786	0.213	<0.001
	weak	0.781	0.219	<0.001
	dead	0.000	0.000	1.000

Transition probability matrix for $t = 1000$ (years).

		to:		
from:		strong	weak	dead
	strong	0.785	0.215	<0.001
	weak	0.785	0.215	<0.001
	dead	0.000	0.000	1.000

Sojourn times (in years).

	estimate	standard error	lower	upper
strong	445	36	379	522
weak	122	10	104	143



Prevalence plots for 'strong' (State 1), 'weak' (State 2), and 'dead' (State 3).

Conclusions

- Multi-state Markov model performs well on this case-study
- Might be a viable alternative for diachronic linguistics
- More information: Van de Velde & De Smet (2021)

References

- Anderwald, L. 2012. 'Variable past tense forms in 19th-century American English: Linking normative grammars and language change'. *American Speech* 87(3): 257-293.
- Baayen, R.H. & P. Milin. 2010. 'Analyzing reaction times'. *International Journal of Psychological Research* 3(2): 12-28.
- Carroll, R., R. Svare & J. Salmons. 2012. 'Quantifying the evolutionary dynamics of German verbs'. *Journal of Historical Linguistics* 2(2): 153-172.
- Cuskley, C.F., 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.
- De Smet, I. 2021. *De sterke werkwoorden in het Nederlands. Een diachroon, kwantitatief onderzoek*. PhD Diss. KU Leuven.
- De Smet, I. & F. Van de Velde. 2020. 'A corpus-based quantitative analysis of twelve centuries of preterite and past participle morphology in Dutch'. *Language Variation and Change* 32: 241-265.
- De Smet, I., & F. Van de Velde. 2019. 'Reassessing the evolution of West-Germanic preterite inflection'. *Diachronica* 36(2): 139-179.
- De Vriendt, S.F.L. 1965. *Sterke werkwoorden en sterke werkwoordsvormen in de 16de eeuw*. Belgisch interuniversitair centrum voor neerlandistiek.
- Jackson, C.H. 2011. 'Multi-state models for panel data: The msm package for R'. *Journal of Statistical Software* 38(8): 1-29.
- Jackson, C.H. 2019. 'Multi-state modelling with R: The msm package'. Version 1.6.8. https://cran.r-project.org/web/packages/msm/vignettes/msm-manual.pdf
- 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(7163): 713-716.
- Van de Velde, F. & I. De Smet. 2021. 'Markov models for multi-state language change'. *Journal of Quantitative Linguistics* (published ahead of print).
- 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, F. & D. Pijpops. Forthc., 2021. 'Investigating lexical effects in syntax with regularized regression (Lasso)'. *Journal of Research Design and Statistics in Linguistics and Communication Science*.
- Winter, B., & Wieling, M. 2016. 'How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling'. *Journal of Language Evolution* 1(1): 7-18.