

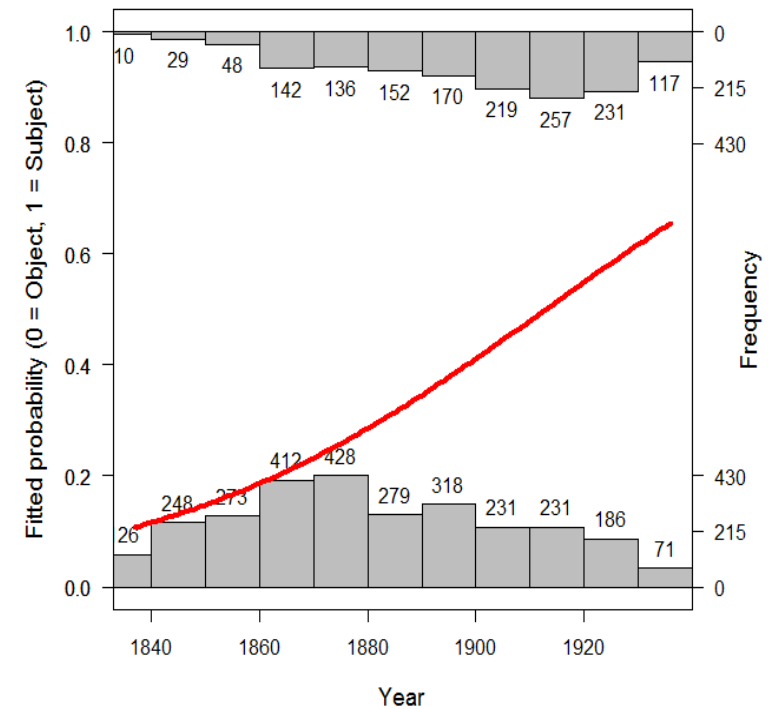
Taming the beast of time

Freek Van de Velde



Status quaestionis

- Historical linguistics has partaken in what Janda (2013) has called the 'quantitative turn' (Hilpert & Gries 2016; Jensen & McGillivray 2017; Van de Velde & Petré 2020)
- This involves the use of TIME as an explanatory variable



What's the problem?

- "historical linguistics (...) can be thought of as the art of making the best use of bad data" (Labov 1994: 11)

- historical linguistics can be thought of as the art of making the best use of bad methods

- Time is an odd beast:

"[T]he fact that time is a dynamic process provides challenges in formulating a model that are not present in settings where a typical linear or logistic regression model might be applied." (Hosmer et al. 2008: 2).

What's the problem?

- Some concrete problems with generalized linear models (Van de Velde & De Smet 2022; Van de Velde, manuscript, for solutions):
 - Non-independence: autocorrelation
 - Generalized linear models assume a smooth monotonic increase/decrease, 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

Methods to consider

- Time series analysis (Koplenig 2017; Koplenig et al. 2016; Van de Velde & Petré 2020: 346-350), including Granger causality (Moscoso del Prado Martín 2014; Rosemeyer & Van de Velde 2020)
- Survival analysis (Van de Velde & Keersmaekers 2020)
- Markov Models (Van de Velde & De Smet 2022)

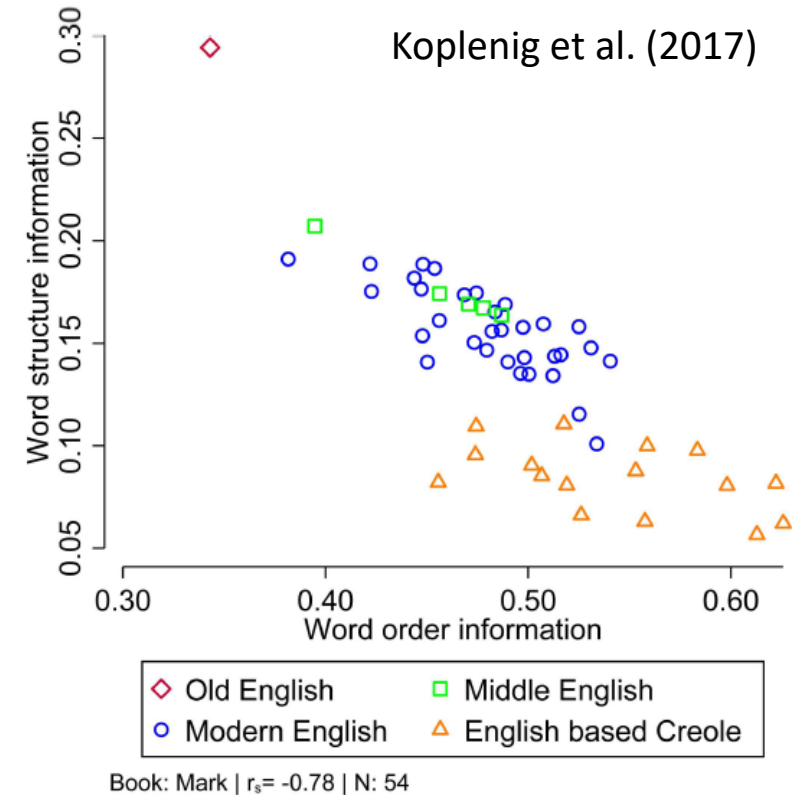
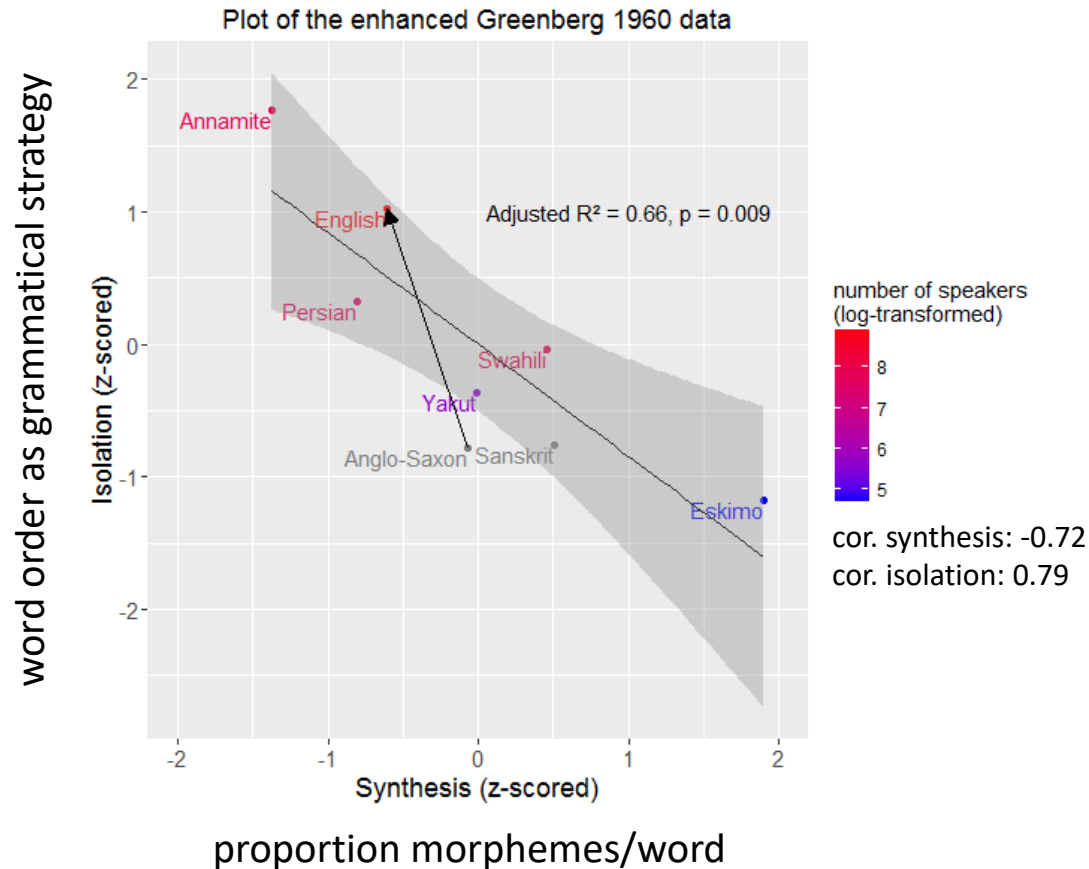
Part 1: Time Series Analysis and Granger Causality

Granger causality

- Correlation does not imply causation
- Historical linguists don't do why questions (Aitchison 2013: 142)
- How can we detect causation?

Correlation and causation

- word order \sim 1 / inflectional morphology



Granger causality

- Technique for comparing two supposedly causally related time series (Granger 1969; Thurman & Fisher 1988)
- General idea: which is better?
 - a) forecast time series A by prior values of A?
 - b) forecast time series A by prior values of A and prior values of B?
 - c) forecast time series B by prior values of B?
 - d) forecast time series B by prior values of B and prior values of A?

} compare } compare
- Example: alcohol intake ~ drunkenness
 - a) forecast drunkenness by prior values of drunkenness? \Rightarrow 😊 because of seasonality and autocorrelation
 - b) forecast drunkenness by prior values of alcohol intake? \Rightarrow 😊
 - c) forecast alcohol intake by prior values of alcohol intake? \Rightarrow 😊
 - d) forecast alcohol intake by prior values of drunkenness? \Rightarrow 😞

Granger causality

- Technique for comparing two supposedly causally related time series (Granger 1969; Thurman & Fisher 1988; Lesmeister 2013)

$$y_t = \beta_0 + \sum_{i=1}^k \beta_i y_{t-i} + \varepsilon_t$$

- General idea: which is better?

$$y_t = \beta_0 + \sum_{i=1}^k \beta_i y_{t-i} + \sum_{i=1}^k \alpha_i x_{t-i} + \varepsilon_t$$

- a) forecast time series A by prior values of A?
 - b) forecast time series A by prior values of A and prior values of B?
 - c) forecast time series B by prior values of B?
 - d) forecast time series B by prior values of B and prior values of A?
- } compare
} compare
} compare

- Example: alcohol intake ~ drunkenness

- a) forecast drunkenness by prior values of drunkenness? \Rightarrow 😊 because of seasonality and autocorrelation
- b) forecast drunkenness by prior values of alcohol intake? \Rightarrow 😊
- c) forecast alcohol intake by prior values of alcohol intake? \Rightarrow 😊
- d) forecast alcohol intake by prior values of drunkenness? \Rightarrow 😞

Granger causality

- Technique for comparing two supposedly causally related time series (Granger 1969; Thurman & Fisher 1988; Lesmeister 2013)

$$y_t = \beta_0 + \sum_{i=1}^k \beta_i y_{t-i} + \varepsilon_t$$

soft ARIMA model
for time series analysis

- General idea: which is better?

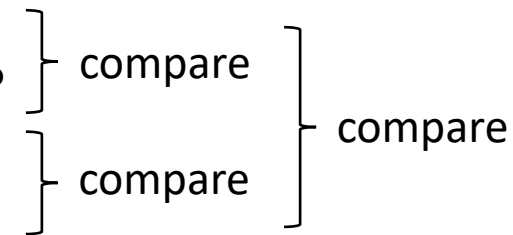
a) forecast time series A by prior values of A?

b) forecast time series A by prior values of A and prior values of B?

c) forecast time series B by prior values of B?

d) forecast time series B by prior values of B and prior values of A?

$$y_t = \beta_0 + \sum_{i=1}^k \beta_i y_{t-i} + \sum_{i=1}^k \alpha_i x_{t-i} + \varepsilon_t$$



- Example: alcohol intake ~ drunkenness

a) forecast drunkenness by prior values of drunkenness? \Rightarrow 😊 because of seasonality and autocorrelation

b) forecast drunkenness by prior values of alcohol intake? \Rightarrow 😊

c) forecast alcohol intake by prior values of alcohol intake? \Rightarrow 😊

d) forecast alcohol intake by prior values of drunkenness? \Rightarrow 😞

Granger causality



- A linguistic case study on Brazilian Portuguese (collaboration with Malte Rosemeyer)

1. *Onde foi você?* [unclefted]
where go.PST.PFV.IND.3SG you
'Where did you go?'

2. *Onde (é) que você foi?* [wh-cleft]
where be.PRS.IND.3SG that you go.PST.PFV.IND.3SG
'Where did you go?'

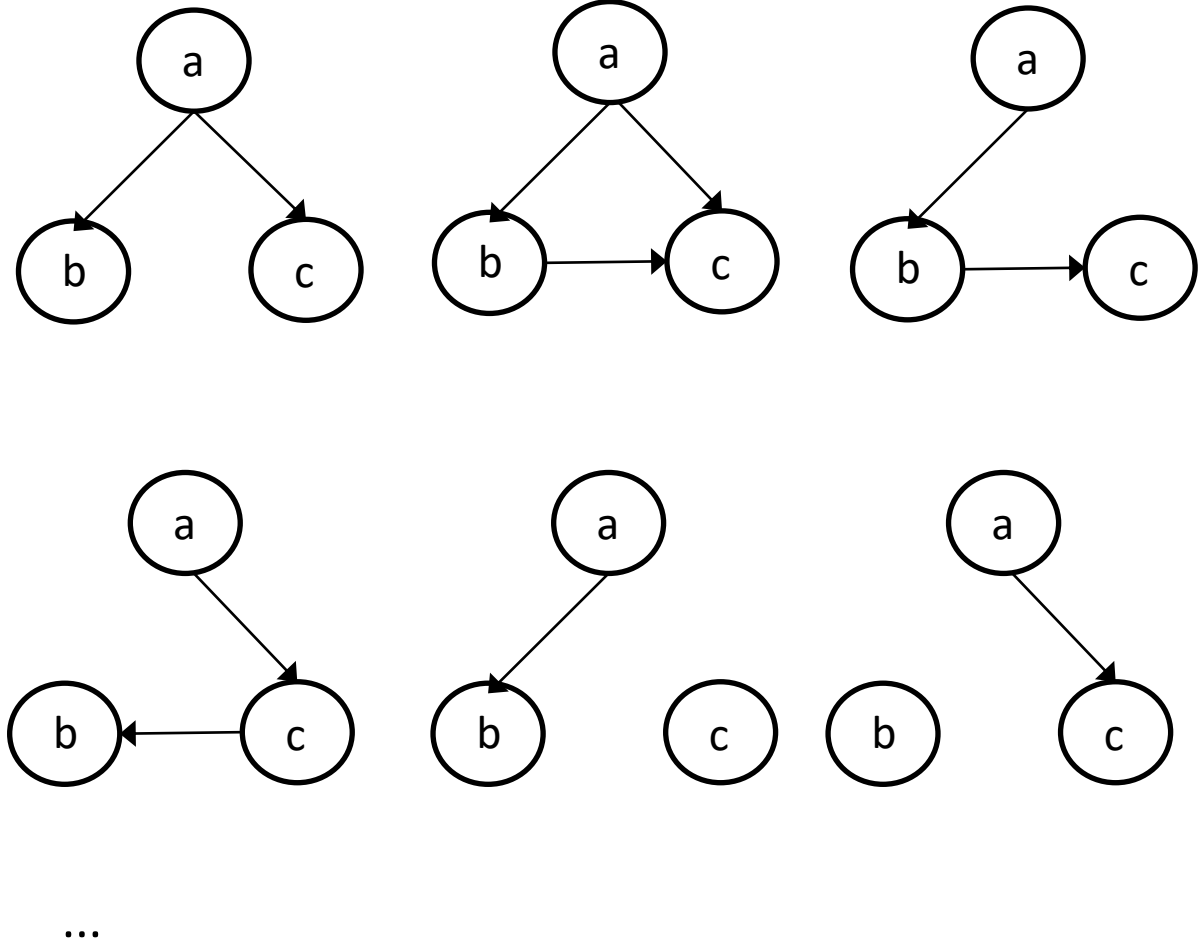
3. *É a Maria que chegou* [that-cleft]
be.PRS.IND.3SG DET.DEF.F Maria that arrive-PST.PFV.IND.3SG
'It is Maria that arrived'

- Rise in SV word order (no more V2-induced inversion) may have triggered the increased use of clefts

Granger causality

- Three time series:
 - a) SV word order
 - b) wh-clefts
 - c) that-clefts

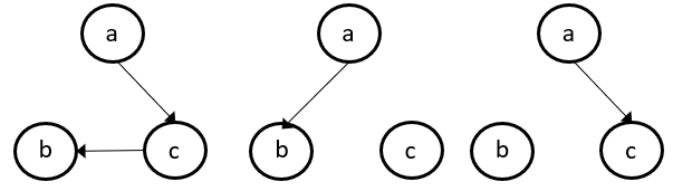
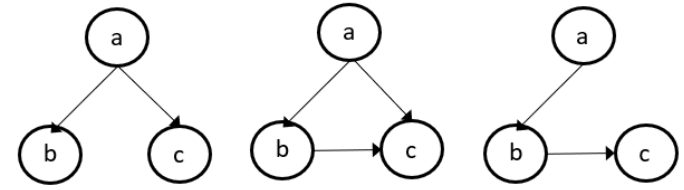
hypotheses:



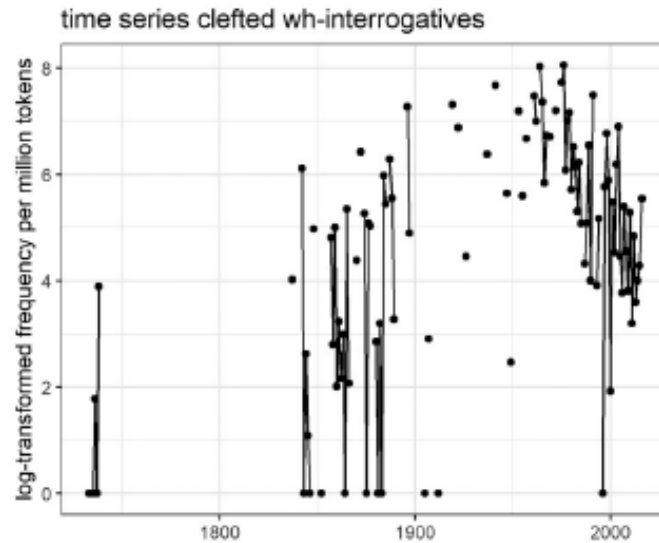
Granger causality

- Three time series:
 - a) SV word order
 - b) wh-clefts
 - c) that-clefts

hypotheses:



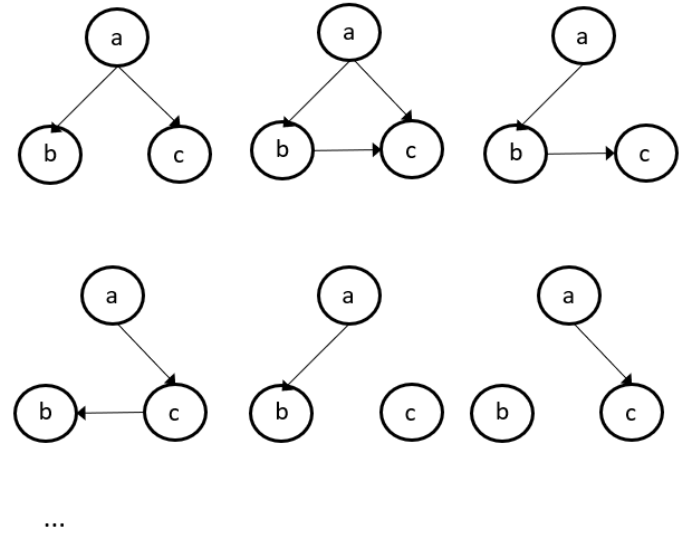
...



Granger causality

- Three time series:
 - a) SV word order
 - b) wh-clefts
 - c) that-clefts

hypotheses:



```
#Granger tests#  
set.order=1
```

```
grangertest(QCY.ts ~ SVY.ts, order=set.order, na.action = na.omit)  
grangertest(SVY.ts ~ QCY.ts, order=set.order, na.action = na.omit)
```

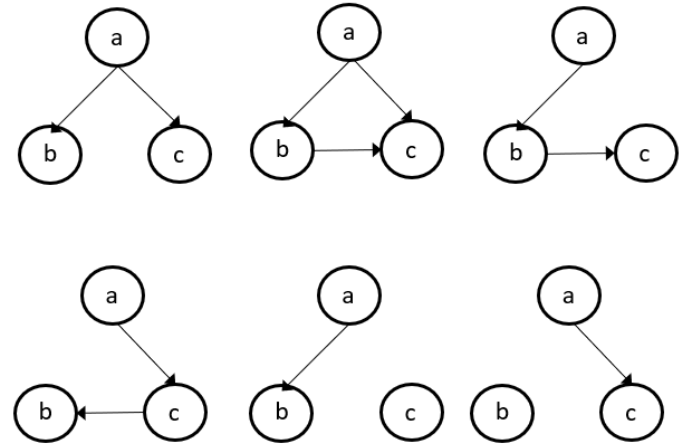
```
grangertest(DCY.ts ~ SVY.ts, order=set.order, na.action = na.omit)  
grangertest(SVY.ts ~ DCY.ts, order=set.order, na.action = na.omit)
```

```
grangertest(DCY.ts ~ QCY.ts, order=set.order, na.action = na.omit)  
grangertest(QCY.ts ~ DCY.ts, order=set.order, na.action = na.omit)
```

Granger causality

- Three time series:
 - a) SV word order
 - b) *wh*-clefts
 - c) *that*-clefts

hypotheses:



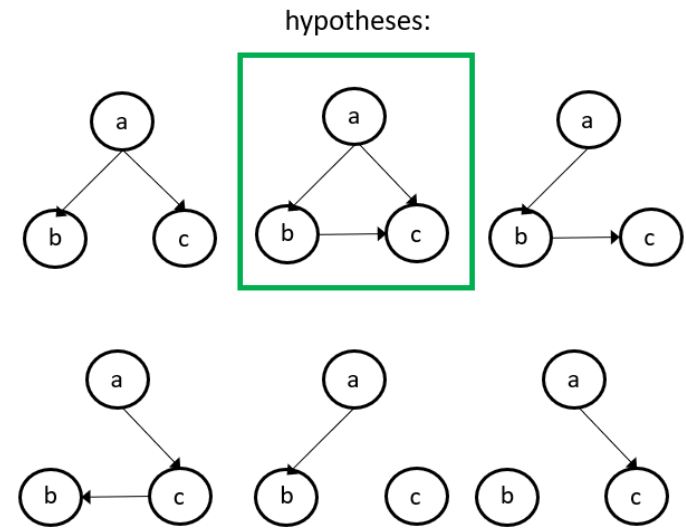
...

Summary of the pairwise Granger Causality tests between the four time series
(light gray shading: $p < 0.05$, dark gray shading: $p < 0.01$)

	SV word order	'That'-clefts	Clefted <i>wh</i> -interrogatives
SV word order causing	–	$p = 0.0410$	$p = 0.0097$
'That'-clefts causing	$p = 0.2872$	–	$p = 0.5007$
Clefted <i>wh</i> -interrogatives causing	$p = 0.1143$	$p = 0.0172$	–

Granger causality

- Three time series:
 - a) SV word order
 - b) *wh*-clefts
 - c) *that*-clefts



Summary of the pairwise Granger Causality tests between the four time series
(light gray shading: $p < 0.05$, dark gray shading: $p < 0.01$)

	SV word order	'That'-clefts	Clefted <i>wh</i> -interrogatives
SV word order causing	–	$p = 0.0410$	$p = 0.0097$
'That'-clefts causing	$p = 0.2872$	–	$p = 0.5007$
Clefted <i>wh</i> -interrogatives causing	$p = 0.1143$	$p = 0.0172$	–

Part 2: Survival Analysis

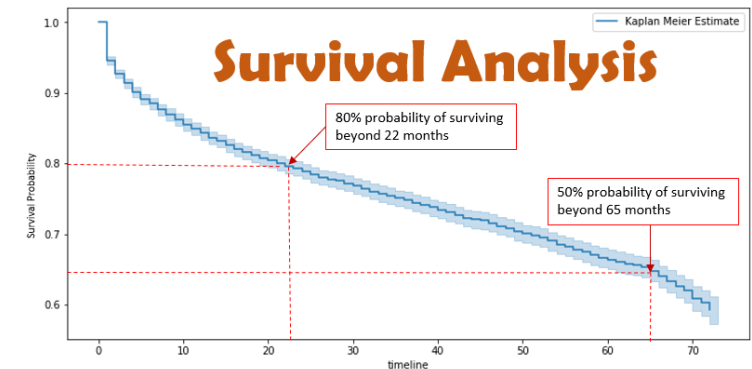
Survival analysis

- Technique to model or predict the time of an 'event', e.g. a patient's death, failure of a machine, winning a Nobel price ...
- Widely applied in different fields, under various names:
 - Medical sciences: 'survival analysis'
 - Engineering: 'reliability analysis' (e.g. infamous bathtub curve)
 - Economics: 'duration analysis'
 - Sociology: 'event history analysis'

not like this:

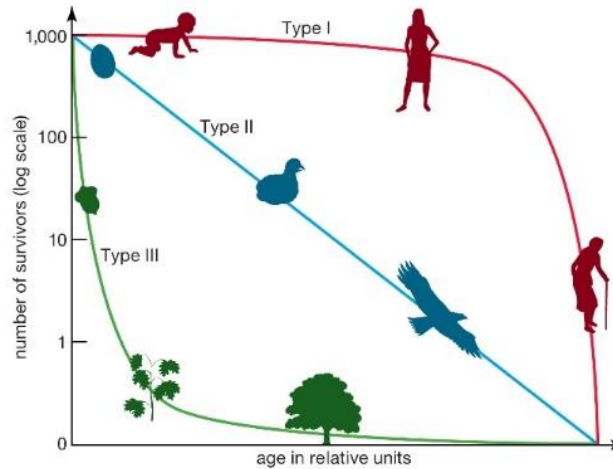


but like this:



Survival analysis

- Different curves:



© 2011 Encyclopædia Britannica, Inc.

- Survival function $S(t)$, with the aid of the Kaplan-Meier estimator (non-parametric) $\hat{S}(t)$

d_j number of dead at time t_j

r_j number of patients at risk at time t_j

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j}\right) & \text{if } t_1 \leq t \end{cases}$$

- More advanced: Cox Proportional Hazard model (parametric)

Case study: survival of words in post-classical Greek

- Details: Van de Velde, Freek & Alek Keersmaekers. 2020. 'What are the determinants of survival curves of words? An evolutionary linguistics approach'. *Evolutionary Linguistic Theory* 2(2): 127-137.
- Dataset of 2217 tokens in Greek papyri from the period 331BC-835AD, almost all of which from Egypt, coded for various variables:



	A	B	C	D	E	F	G	H
1	LEMMA	POS	POS_SPEC	FREQ	YEAR_FIRST	YEAR_LAST	WORDS_SUBCORPUS	PHON_SIZE
2	ἕξ	a	cardinal	277	-257	710	997702	2
3	ὥνή	n	lexicaal	279	-259	352	903235	3
4	ἐπιτηρέω	v	lexicaal	10	-161	352	740334	8
5	ἐσπέρα	n	lexicaal	12	136	559	518126	6
6	δαύζω	v	lexicaal	10	-18	752	731920	5
7	ἀνάκρισις	n	lexicaal	11	-241	270	689209	9
8	προσβολή	n	lexicaal	13	-260	375	919692	8
9	δάνειον	n	lexicaal	205	-257	570	957183	7
10	συμβολαιογράφος	n	lexicaal	12	452	719	83961	15
11	κεντηνάριος	n	lexicaal	11	315	710	168216	11
12	ἔνγραπτος	a	lexicaal	23	-221	254	627458	9
13	ἀποχωρέω	v	lexicaal	13	-258	180	582216	8
14	,3	a	numeral_symbol	14	-244	248	649742	2
15	ἱκεσία	n	lexicaal	12	317	570	122967	6
16	διοικέω	v	lexicaal	100	-263	697	993049	7
17	σξα	a	numeral_symbol	11	-245	705	927667	3
18	φόρος	n	lexicaal	284	-259	641	984532	5
19	εὐκαιρος	d	lexicaal	17	-257	311	822575	8
20	ἐπιφανής	a	lexicaal	320	-212	392	796410	8

Case study: survival of words in post-classical Greek

- Details: Van de Velde, Freek & Alek Keersmaekers. 2020. 'What are the determinants of survival curves of words? An evolutionary linguistics approach'. *Evolutionary Linguistic Theory* 2(2): 127-137.
- Dataset of 2217 tokens in Greek papyri from the period 331BC-835AD, almost all of which from Egypt, coded for various variables:

Variable	Levels or range	Type frequency
Size (in characters)	Numeric, range [1; 17]	2217
Size (3 equally spaced groups)	Small	1105
	Medium	1077
	Large	35
First attestation	Numeric, range [331BC; 3BC]	2217
Last attestation	Numeric, range [235BC; 700AD]	2217
Part-of-speech	Noun	1019
	Verb	863
	Adjective	335
Frequency (transformed)	Numeric, range [-2.00; 1.15]	2217
Frequency (3 equally spaced groups)	Low	1732
	Medium	451
	High	34
Observed span (in years)	Numeric, range [15; 1056]	2217

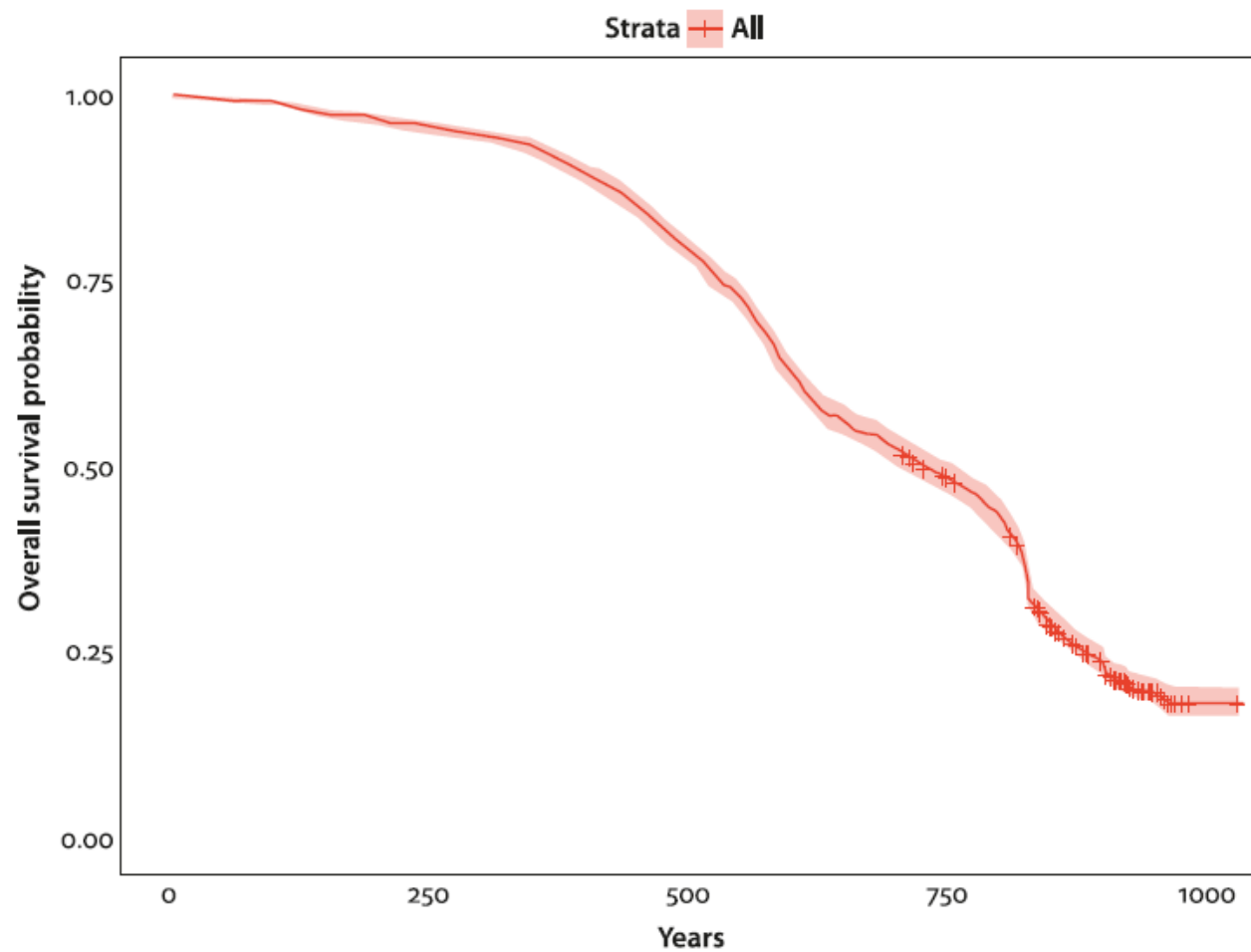
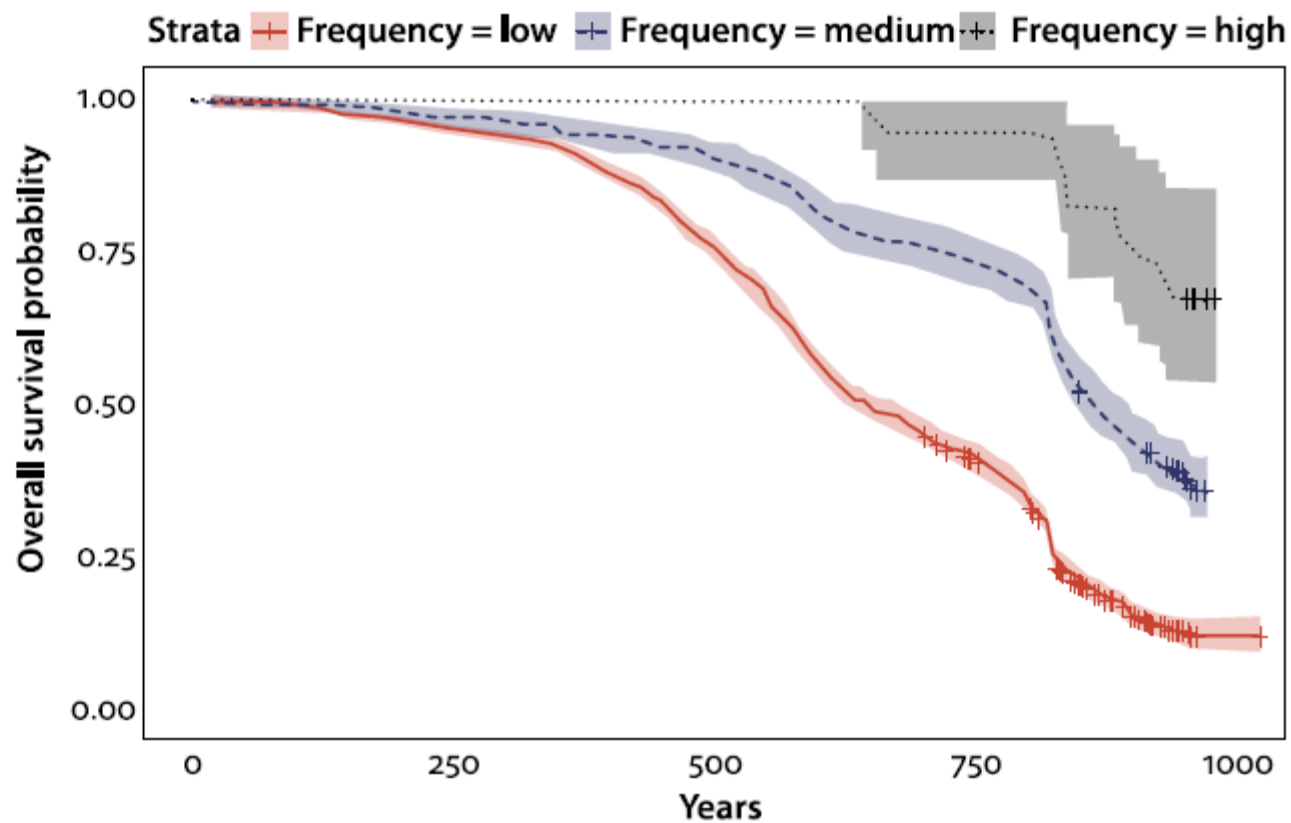


Figure 1. Kaplan-Meier Curve for survival of all types ($n=2217$). The x-axis gives the time span in years over which a lexeme has been observed. It is not the date of attestation. ‘Censored’ types (words that are still alive in 700AD) are indicated as crosses (+)

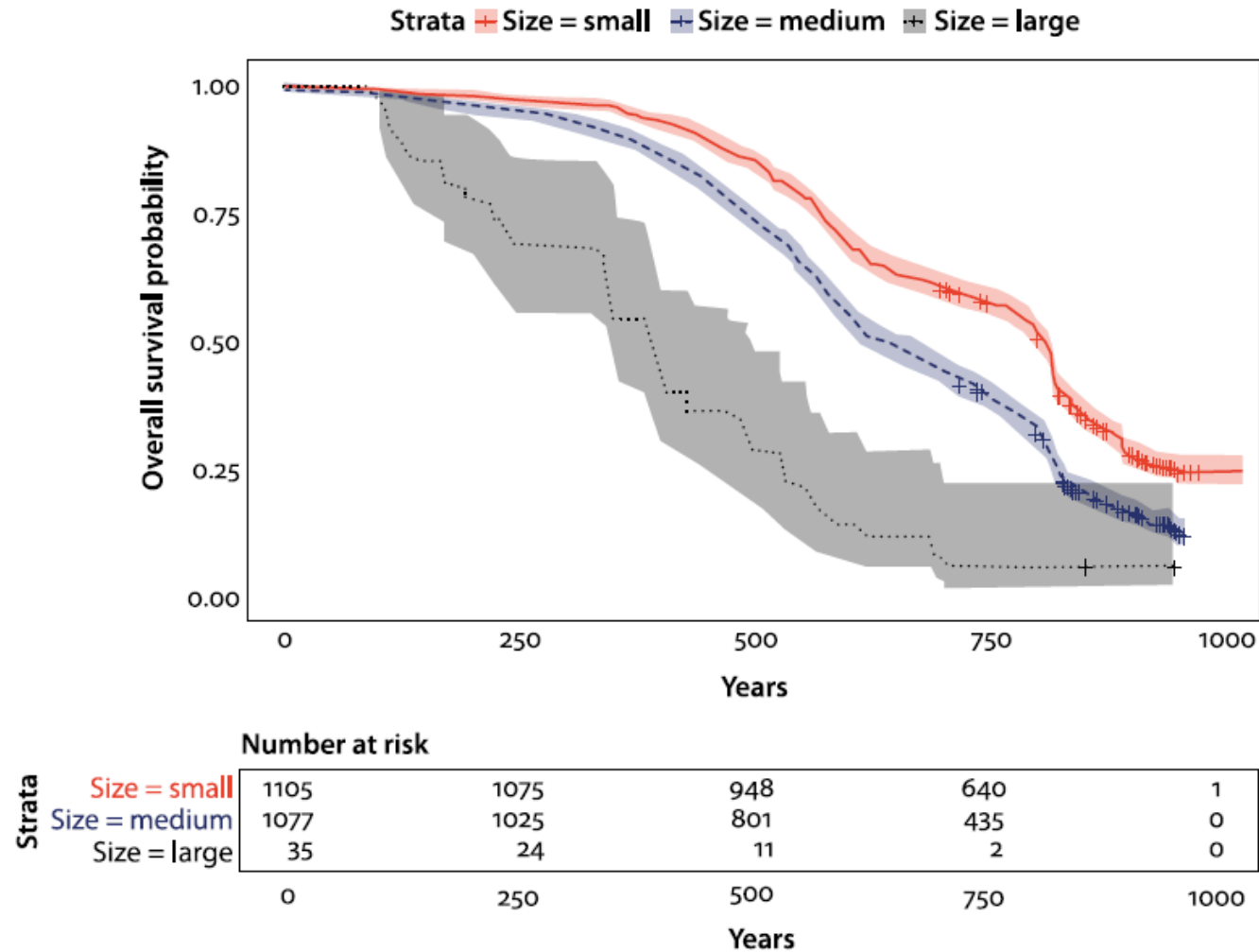


Difference between strata is significant (log-rank test), $p < 0.001$

Number at risk

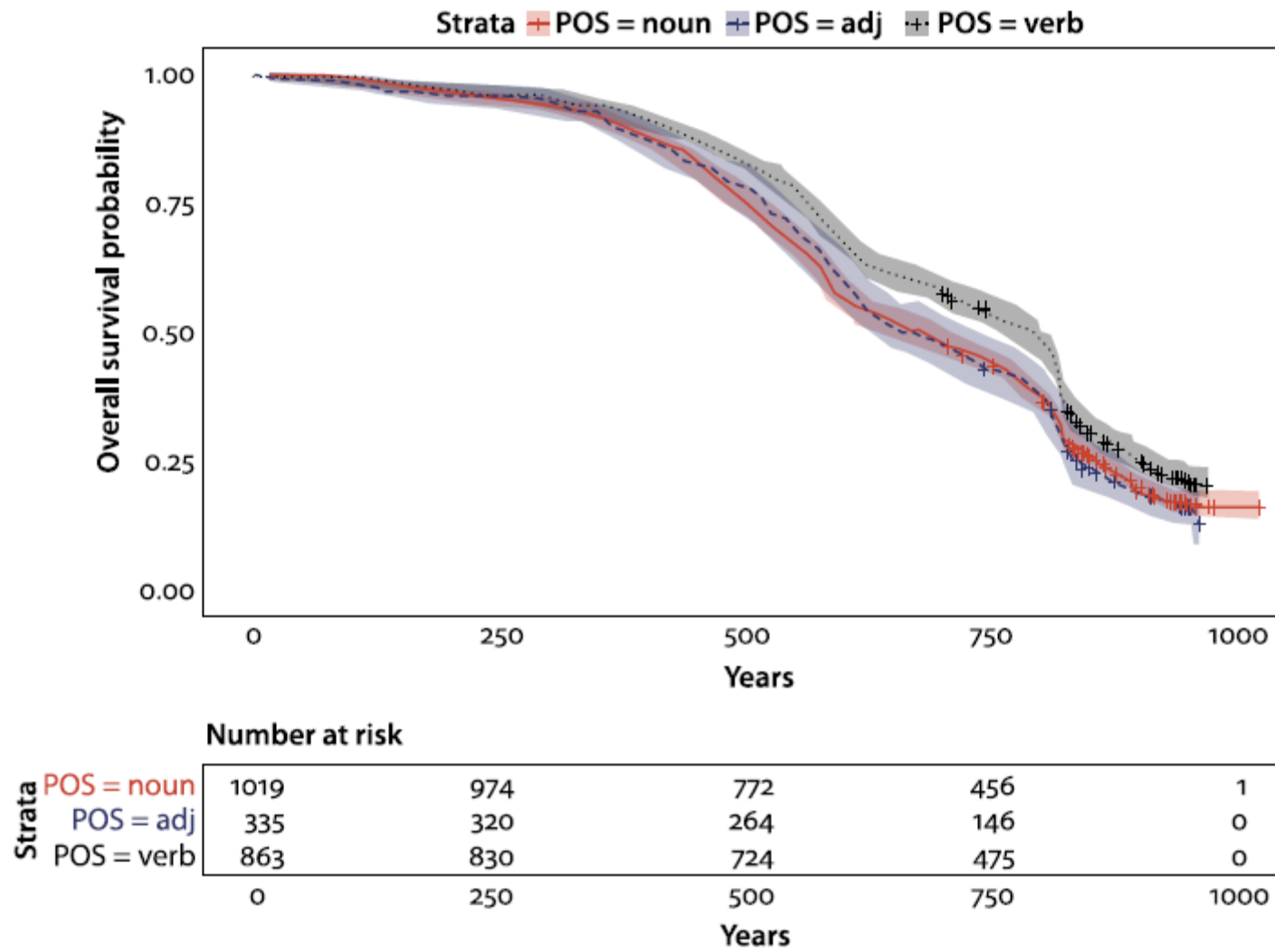
Strata	0	250	500	750	1000
Frequency = low	1732	1652	1318	714	1
Frequency = medium	451	438	408	331	0
Frequency = high	34	34	34	32	0

Years



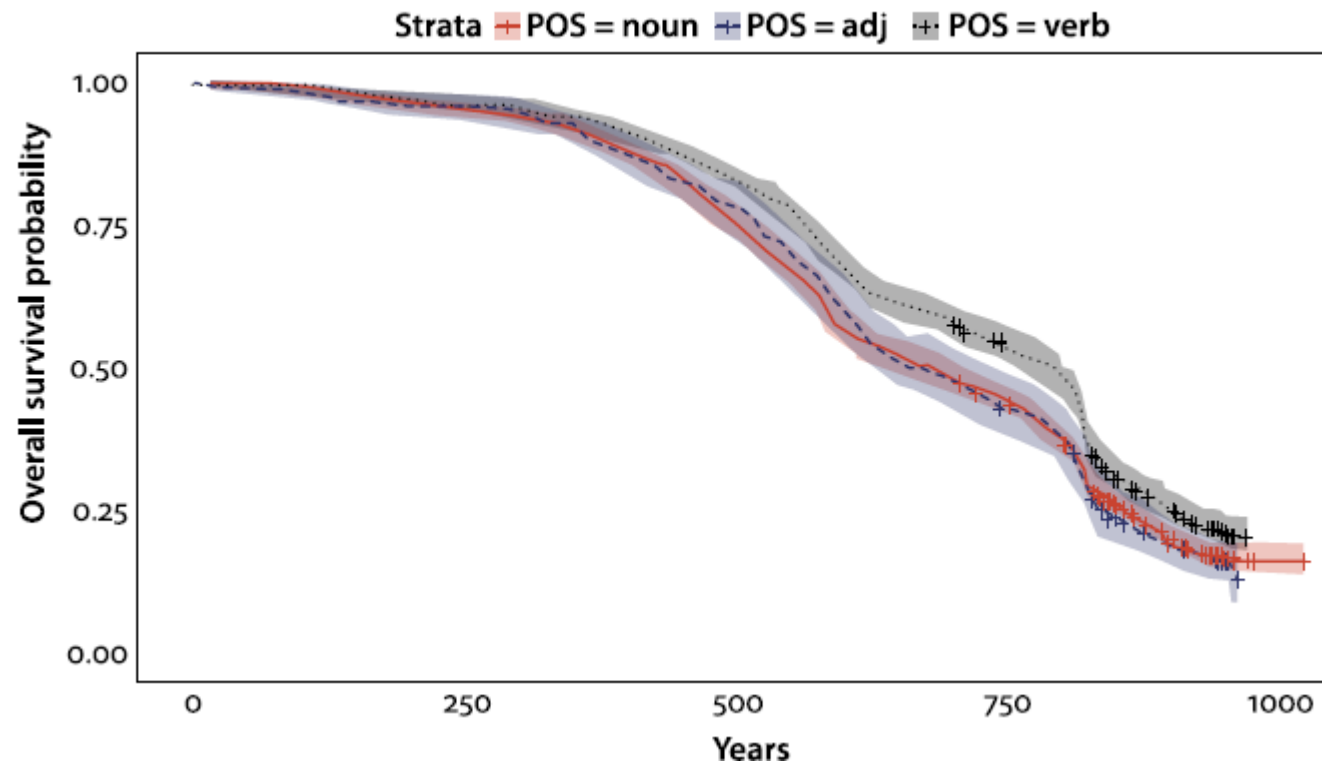
Difference between strata is significant (log-rank test), $p < 0.001$

Figure 3. Kaplan-Meier Curve for survival of PHONETIC SIZE strata. The x-axis gives the time span in years over which a lexeme has been observed. 'Censored' types are indicated as crosses (+)



Difference between strata is significant (log-rank test), $p < 0.001$

Figure 4. Kaplan-Meier Curve for survival of PART-OF-SPEECH strata. The x-axis gives the time span in years over which a lexeme has been observed. ‘Censored’ types are indicated as crosses (+)



⇔ Pagel et al. (2007): $N > V > A$

⇒ Grossman & Polis (2017)

Different strategies: V are innovative because they are derivationally mutilated or undergo polysemous radiation

Extra

- Parametric approach: Cox Proportional Hazard model
- Hazard function:

$$\hat{\lambda}(t_j) = \frac{d_j}{r_j}$$

$$\lambda(t; Z) = \lambda_0(t) e^{(\sum_i^n \beta_i Z_i)}$$

car still at risk in the denominator:



Cox Proportional Hazard Model

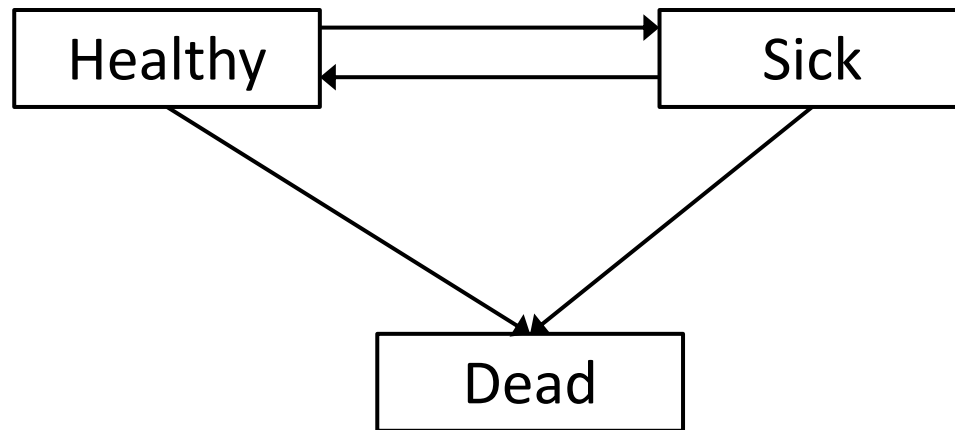
Variable		Coefficient	z-score	p-value
Frequency	Numeric	-0.717	-13.135	< 0.0001
Phonetic size	Numeric	0.123	10.443	< 0.0001
Part-of-speech	Noun	(reference level)	-	-
	Adjective	-0.118	-1.696	0.0899
	Verb	-0.281	-5.316	< 0.0001
First attestation	Numeric	0.005	10.574	< 0.0001

Part 3: Multi-State Markov Model

Multi-State Markov Model

- Belongs to the family of Survival Analysis
 - 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 on Dutch preterite formation

JOURNAL OF QUANTITATIVE LINGUISTICS
2022, VOL. 29, NO. 3, 314–338
<https://doi.org/10.1080/09296174.2021.1877004>

 **Routledge**
Taylor & Francis Group

 OPEN ACCESS  Check for updates

Markov Models for Multi-state Language Change

Freek Van de Velde ^a and Isabeau De Smet ^{a,b}

^aDepartment of Linguistics, KU Leuven, Leuven, Belgium; ^bFWO Vlaanderen, Bruxelles, Belgium



Case-study: Germanic preterites (Dutch)

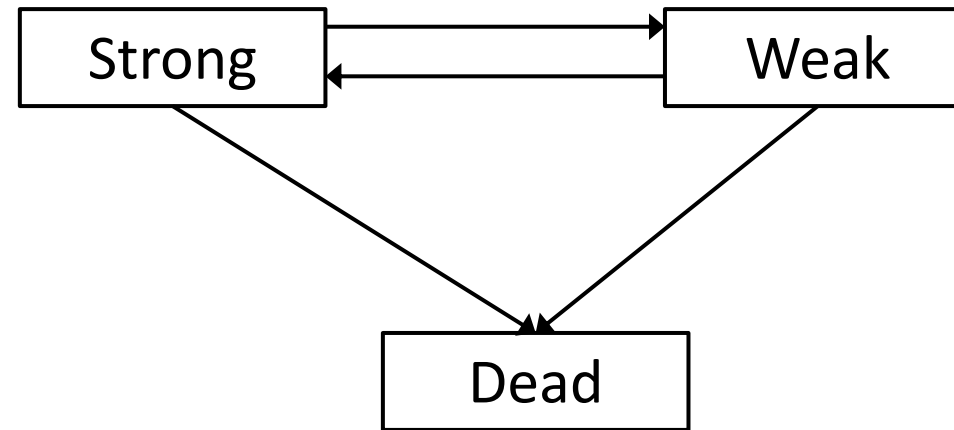
- Germanic languages have two morphological strategies for building preterites (not counting analytic perfects, *he has written a book*):
 1. Strong inflection
 - English *sing – sang*
 - Ablaut, based on Indo-European aspectual system (perfect > preterite)
 2. Weak inflection
 - English *work – worked*
 - Dental suffix, based on an analytic formation [VERB + **d^heh*₁-, **d^hoh*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'



Multi-state Markov Model

$$q_{rs}(t, Z(t)) = \lim_{\Delta t \rightarrow 0} \frac{P(S(t + \Delta t) = s | S(t) = r)}{\Delta t}$$

$$Q = \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{pmatrix} = \begin{pmatrix} -(q_{12} + q_{13}) & q_{12} & q_{13} \\ q_{21} & -(q_{21} + q_{23}) & q_{23} \\ 0 & 0 & -(q_{31} + q_{32}) = 0 \end{pmatrix}$$

$$P(t) = e^{tQ}$$

$$q_{rs}(Z(t)) = q_{rs}^{\text{baseline}} e^{\beta_{rs} Z(t)}$$

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
- 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

Multi-state Markov Model, for Dutch preterites

$$q_{rs}(z(t)) = q_{rs}^{baseline} e^{(\beta_{rs}^{mf} Z_{mf}(t) + \beta_{rs}^{hf} Z_{hf}(t) + \beta_{rs}^{ABC} Z_{ABC}(t) + \beta_{rs}^{ABA} Z_{ABA}(t))}$$

(mf = mid-frequency, hf: high-frequency, ABC: vowel pattern ABC, ABA: vowel pattern ABA)

Results

covariate		transition	coefficient	confidence interval
FREQUENCY	low	(reference level)		
	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).

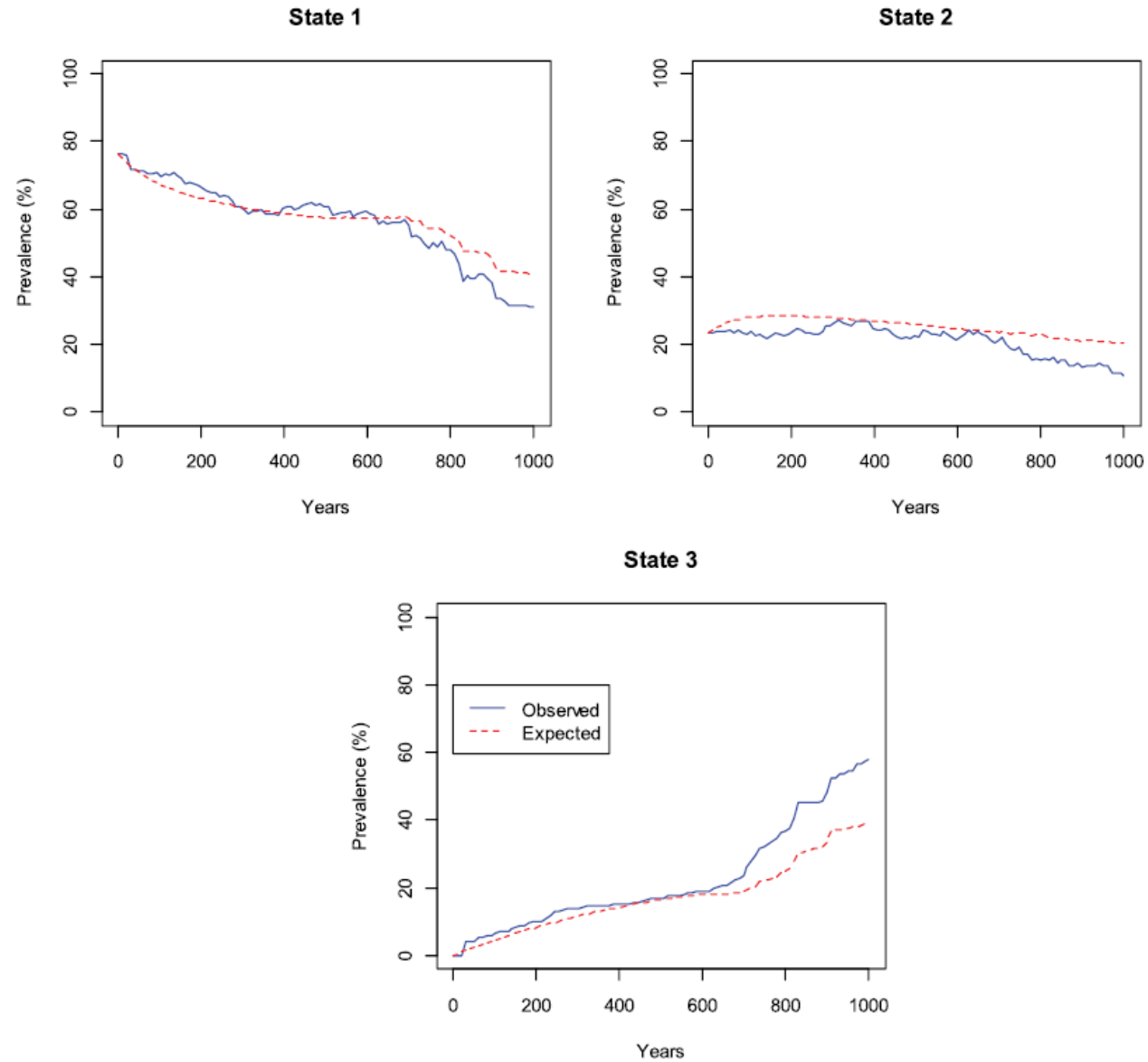
		to:		
from:		strong	weak	dead
	strong	0.786	0.213	<0.001
	weak	0.781	0.219	<0.001
	dead	0.000	0.000	1.000

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).

References

- Aitchison, J. 2013. *Language change. Progress or decay?* 4th edn. Cambridge: Cambridge University Press
- Anderwald, L. 2012. 'Variable past tense forms in 19th-century American English: Linking normative grammars and language change'. *American Speech* 87(3): 257-293
- 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.
- Granger, C.W. 1969. 'Investigating causal relations by econometric models and cross-spectral methods'. *Econometrica* 37: 424-438.
- Grossman, E. & S. Polis. 2017. 'Overall borrowing and borrowing in basic vocabulary: a typological perspective on lexical change in Ancient Egyptian-Coptic'. *Paper presented at 50th Annual Meeting of the Societas Linguistica Europaea (SLE 2017)*, Zurich.
- Hilpert, M. & S.T. Gries. 2016. 'Quantitative approaches to diachronic corpus linguistics'. In: M. Kytö & P. Pahta (eds.), *The Cambridge handbook of English historical linguistics*. Cambridge: Cambridge University Press. 36-53.
- Hosmer, D.W., S. Lemeshow & S. May. 2008. *Applied survival analysis: regression modeling of time-to-event data*. 2nd edn. Hoboken: Wiley.
- Jensen, G.B. & B. McGillivray. 2017. *Quantitative historical linguistics. A corpus framework*. Oxford: Oxford University Press.
- 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.
- Koplenig, A., P. Meyer, S. Wolfer & C. Müller-Spitzer. 2017. 'The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort'. *PLoS One* 12: e0173614
- Labov, W. 1994. *Principles of linguistic change, vol. 1: internal factors*. Oxford: Blackwell.
- Lieberman, E., J.-B. Michel, J. Jackson, T. Tang & M.A. Nowak. 2007. 'Quantifying the evolutionary dynamics of language'. *Nature* 449(7163): 713-716.

References

- Moscoso del Prado Martín, Fermín. 2014. 'Grammatical change begins within the word: causal modeling of the co-evolution of Icelandic morphology and syntax'. *Proceedings of the Annual Meeting of the Cognitive Science Society* 36: 2657-2662.
- Pagel, M., Q.D. Atkinson & A. Meade. 2007. 'Frequency of word-use predicts rates of lexical evolution throughout Indo-European history'. *Nature* 449(7163): 717-720.
- Rosemeyer, M. & F. Van de Velde. 2021. 'On cause and correlation in language change. Word order and clefting in Brazilian Portuguese'. *Language Dynamics and Change* 11(1): 130-166.
- Thurman, W.N. & M.E. Fisher. 1988. 'Chickens, eggs, and causality, or which came first?' *American Journal of Agricultural Economics* 70(2): 237-238.
- Van de Velde, F. & I. De Smet. 2022. 'Markov models for multi-state language change'. *Journal of Quantitative Linguistics* 29(3): 314-338.
- Van de Velde, F. Manuscript. 'A few comments on the use of the time variable in historical corpus studies'.
- Van de Velde, F. & A. Keersmaekers. 2020. 'What are the determinants of survival curves of words? An evolutionary linguistics approach'. *Evolutionary Linguistic Theory* 2(2): 127-137.
- 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.

Contact me for:

- Questions
- Help with your own research
- Full R code
- Helpful bibliographic references
- Exchange of ideas
- ...

freek.vandeveld@kuleuven.be

(Or google me)