Topical advection as a baseline model for corpus-based lexical dynamics

Andres Karjus[1], **Richard A. Blythe**[1,2], **Simon Kirby**[1], **Kenny Smith**[1] | [1]Centre for Language Evolution, [2]School of Physics and Astronomy; **University of Edinburgh**

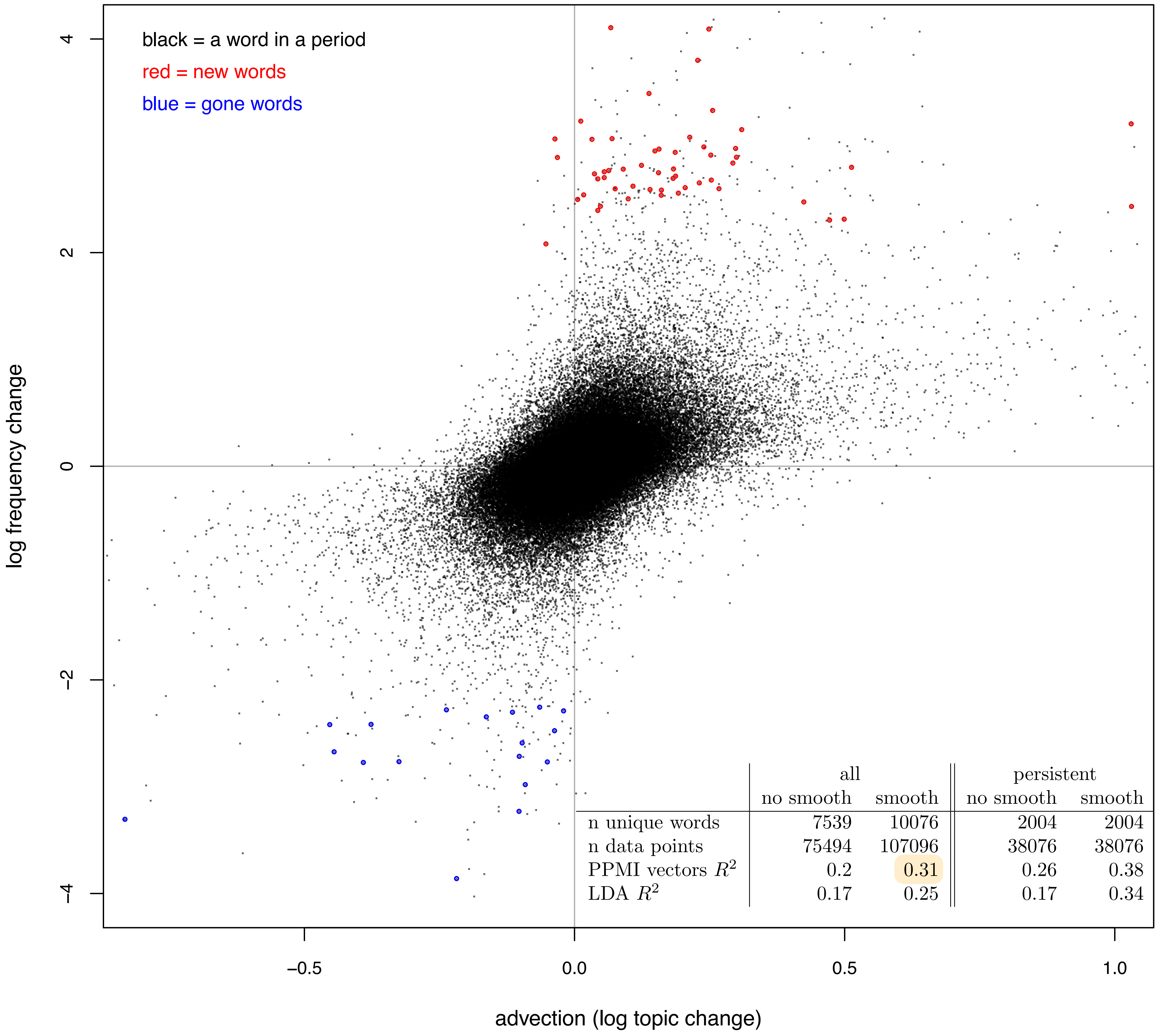
Supported by: European Union European Regional Development Fund; Investing in your future

Overview: word frequencies and topic frequencies

1. Much of research in corpus-based evolutionary language dynamics research (and corpus linguistics in general) relies on token frequencies of linguistic elements in texts.
2. Frequency used as a proxy for the popularity or selective fitness of an element, and as an explanatory factor in quantitative analyses of various linguistic processes.
3. However, a number of recent works: corpus frequencies may be misleading. Attributed to potentially unbalanced sampling of genres and registers, but also: corpora are composed of contemporary media and fiction texts, their underlying topics being reflective of current cultural and socio-political trends.
4. Solution: control for diachronic topical fluctuations by quantifying the frequency change of a word's topic.
5. *advection*: 'the transport of substance, particularly fluids, by bulk motion' (analogy: words being carried along by their topics). Formalized as the weighted mean of the log frequency changes of the (top) context/topic words of the target word.

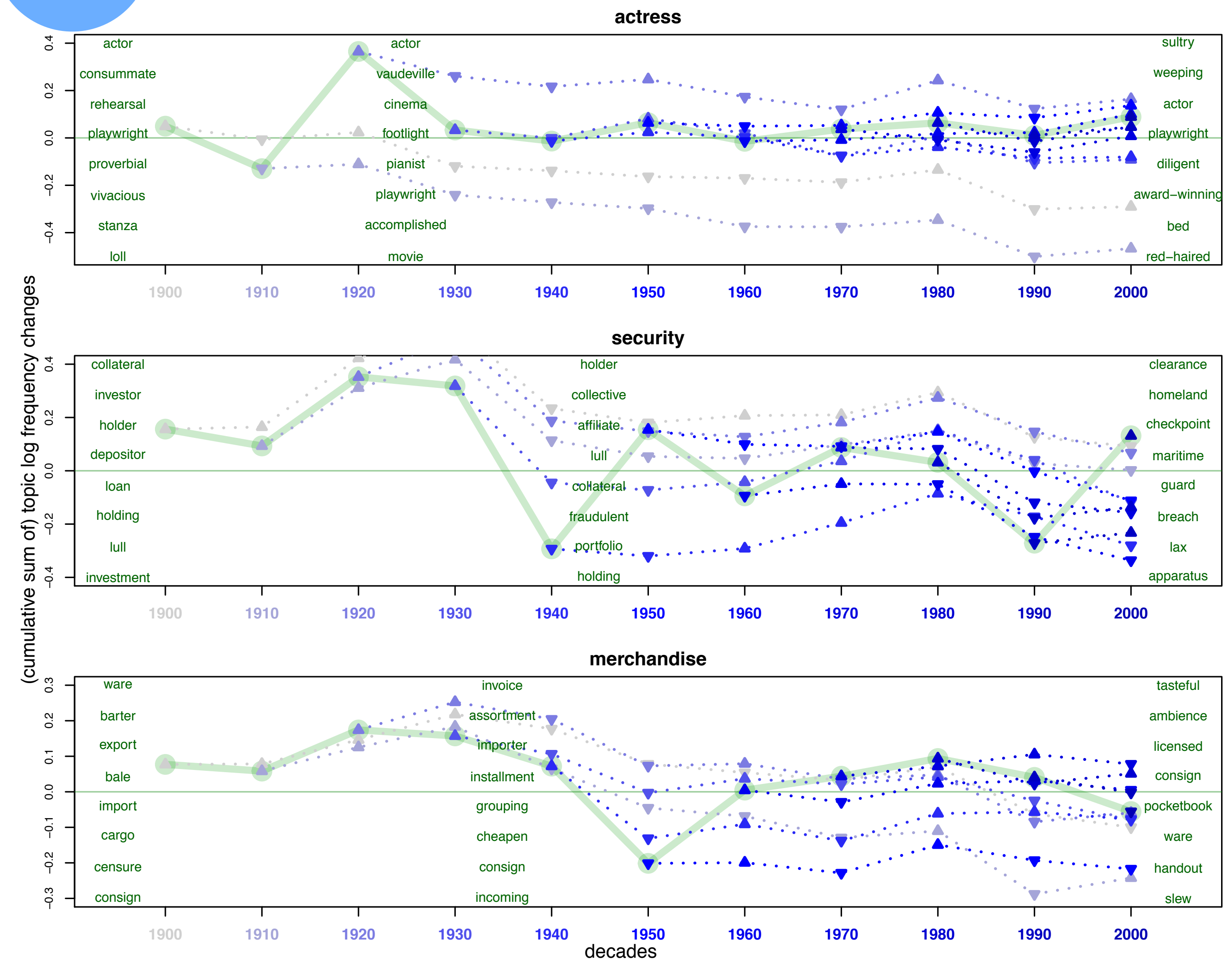
$$advection(w_t) := wMean\big(\{logChange(N_{i_t}) \mid i = 1, ...m\}, W_{1:m}\big) \quad logChange(w_t) := log(w_{freq_t} + 1) - log(w_{freq_{t-1}} + 1)$$

Main Result: advection describes 20-30% of the variability in word frequency changes (nouns in the Corpus of Historical American English, decades 1810-2000)



	all		persistent	
	no smooth	smooth	no smooth	smooth
n unique words	7539	10076	2004	2004
n data points	75494	107096	38076	38076
PPMI vectors R^2	0.2	0.31	0.26	0.38
LDA R^2	0.17	0.25	0.17	0.34

Tracing frequency change of topics over time



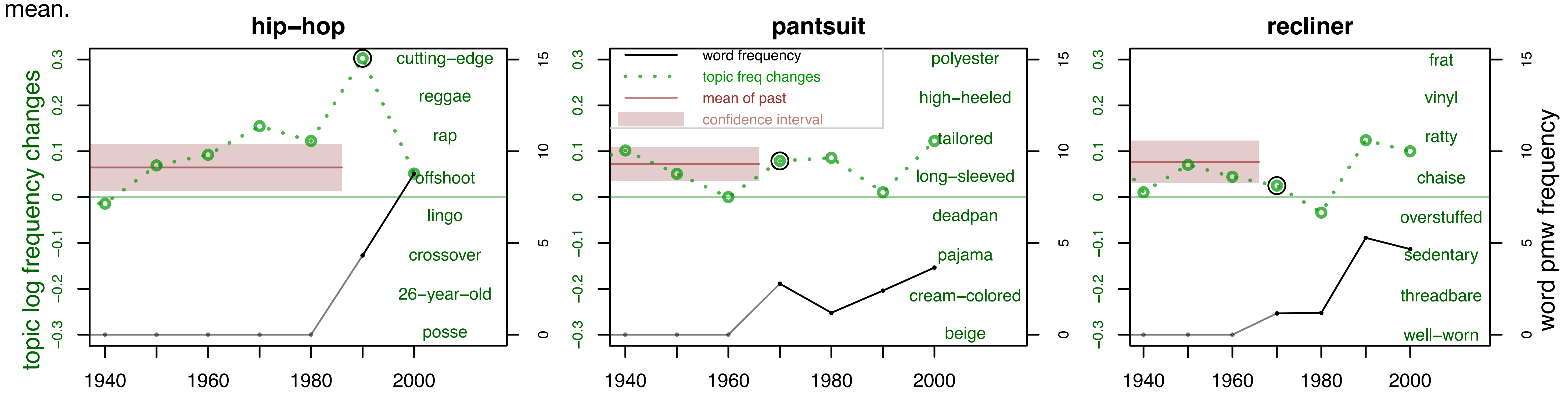
Simulated change in an artificial corpus

We simulated artificial language change using two subcorpora from the synchronic Corpus of Contemporary American English, 'academic' and 'spoken'. We defined the former as the first 'period' and the latter as the later, simulating a scenario where a language changes from academic to spoken in style and content. As before, we calculated the advection values of nouns, which again correlate with frequency change: R^2=0.52 without and R^2=0.8 with smoothing (smoothing here refers to concatenating two periods for the purposes of evaluating the topics).

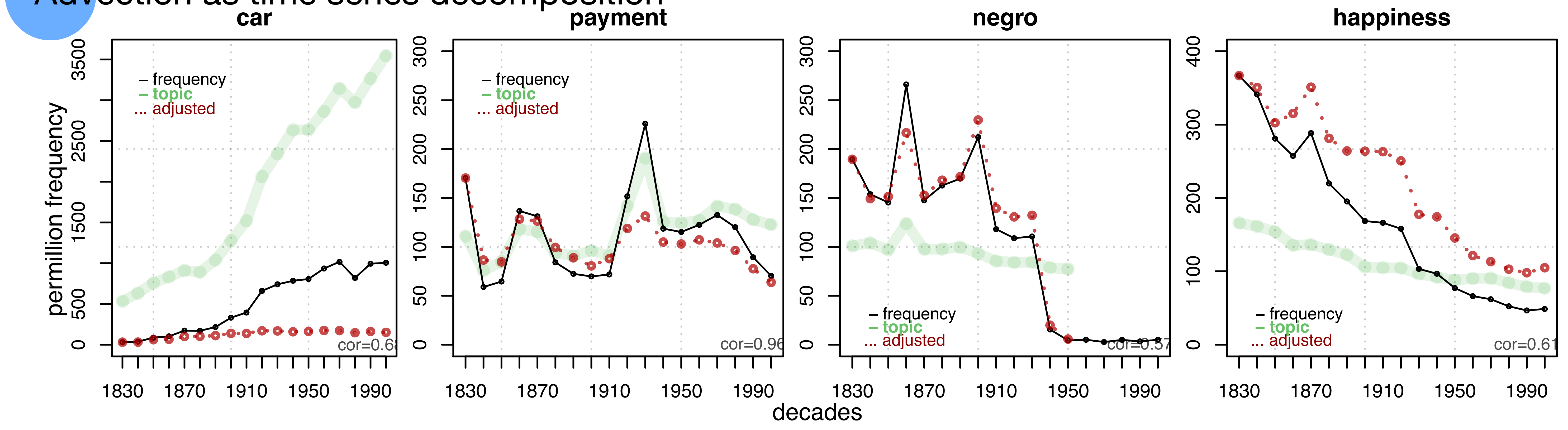
top decreased	freq change	advection	top increased	freq change	advection
supra	-4.95	-1.29	*sir*	+3.87	+1.03
subscale	-4.75	-1.82	*guy*	+3.74	+1.23
coefficient	-4.61	-1.83	*mom*	+3.58	+1.11
variable	-4.42	-1.95	*ma'am*	+3.53	+0.84
variance	-4.1	-1.67	*recount*	+3.08	+0.15
self-efficacy	-4.08	-1.46	*dad*	+3.06	+0.93
regression	-4.03	-1.88	*cop*	+3.05	+0.77
respondent	-4.01	-1.08	*lot*	+3.02	+1.01
learner	-3.93	-1.29	*heck*	+2.95	+0.51
preservice	-3.85	-1.37	*correspondent*	+2.93	+0.76

Advection also predicts lexical innovation

The rise of a topic tends to give rise to new words (presumably to meet growing communicative needs). Out of 133 successful novel nouns that came about in the 1970s-2000s, 55% have an advection value at the time of entry that is above the (95% confidence interval of the) mean of the frequency changes of its topic in the past 100 years; 38% are around the mean, 7% below the (lower bound of the confidence interval of the) mean.



Advection as time series decomposition


car
payment
negro
happiness
– frequency
– topic
... adjusted
permillion frequency
decades
cor=0.6
cor=0.96
cor=0.57
cor=0.61