Modelling lexical interactions in diachronic corpora

Andres Karjus, Richard A. Blythe, Simon Kirby, Kenny Smith

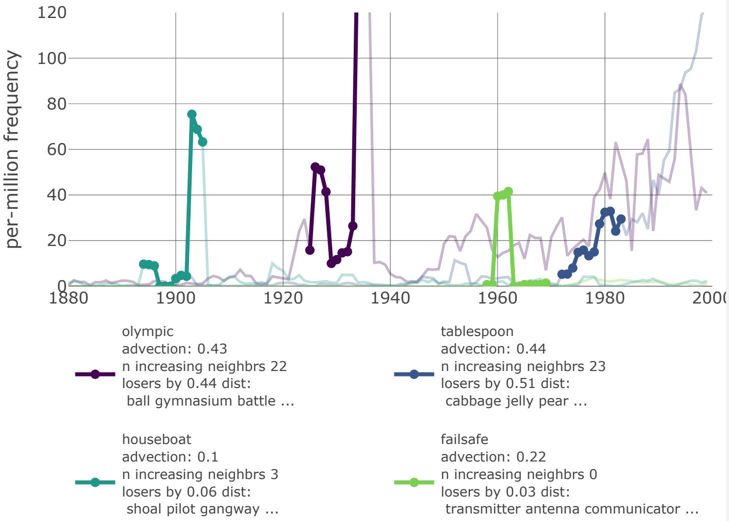
Centre for Language Evolution, University of Edinburgh

Introduction

- Hypothesis: frequency change in a word will lead to direct competition with (and possibly replacement of) near-synonym(s), unless the lexical subspace experiences high communicative need.
- Is it possible to describe some variance in terms of which successful words compete with their neighbors and which do not?

Data

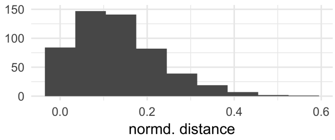
522 unique words (COHA, 1890-1999) with frequency increase $\ln \geq 2$ between any 2 successive spans of 10 years (& occur in ≥ 2 years & ≥ 100 times in the latter span).

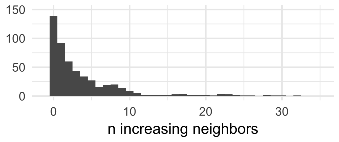


Quantifying competition

- Embed targets into vector space (LSA) of the preceding decade, compute semantic neighbors
- Important: word occurrence probabilities sum up to 1; increase in x => decrease in y.
- The measure: where probability mass gets equalized, i.e., target increase$\geq \sum$(neighbors' decreases). Either cosine distance, or n increasing neighbors.
- Indicates if the increasing target replaced semantically close word(s) (direct competition, obvious likely source of probability mass).
- Example: *relativism*, increasing +13.2pmw between 1965-1974, 1975-1984:

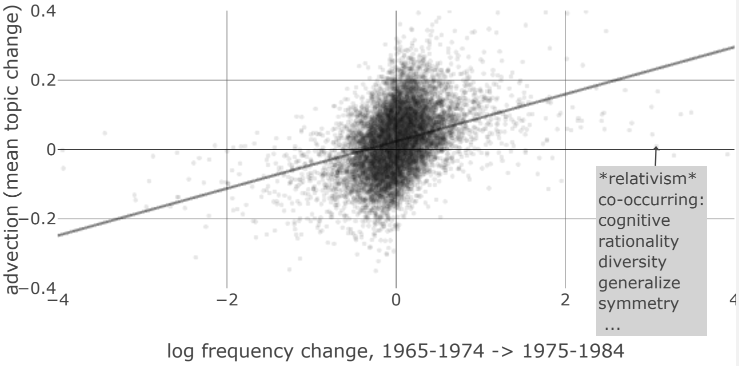
word	freq. change	cumulative sum of decreases	cosine sim	normd. dist
relativism	**+13.2**			
marxism	-5.68	5.68	0.68	0
thesis	+9.00	5.68	0.67	0.01
jacksonian	-11.64	17.32>**13.2**	0.66	**0.03**



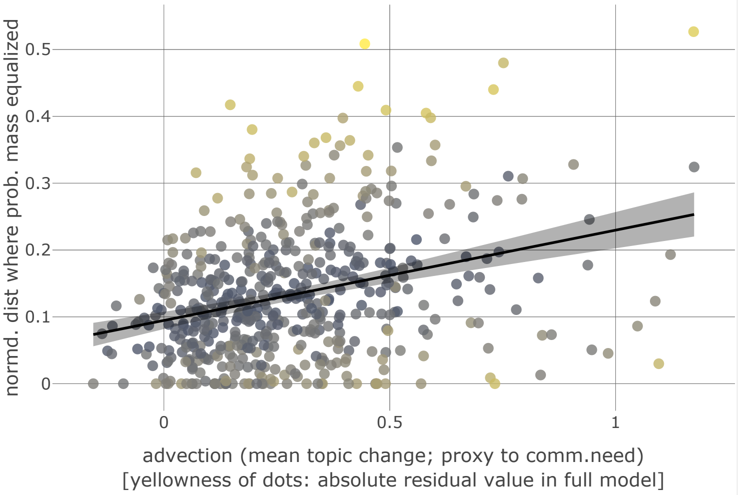


Communicative need

Topical advection as a proxy: weighted mean log frequency change in the top n (PPMI-weighted) context words of the target.



Results



Linear regression model predicting the cosine distance (normalized by value of top neighbor) where probability mass gets equalized

	Estimate	p	clearer competition signal if...
advection	0.1157	<0.001	lower comm. need
closest sem neighbor	0.2519	<0.001	dense subspace
occurs in n years	0.0087	<0.001	bursty series
abs. freq. change	0.0005	0.003	lower freq (change)
max %decrease	0.0009	<0.001	a clear loser

R^2=0.24, F=13.32(13,508), p<0.001

Also controlled for in the model, but all p>0.05: • standard deviation of yearly frequencies (burstiness) • semantic subspace instability • uniqueness of the form • smallest edit distance among closest semantic neighbors • polsemy • leftover probability mass • age of the word in the corpus • target decade.

Conclusions

Controlling for a range of factors, communicative need (operationalized by advection), describes a moderate amount of variance in competitive interactions between words: low advection words are more likely to replace a word with a similar meaning. Presumably high comm.need facilitates the co-existence of similar words.

Interactive poster with appendix: http://andreskarjus.github.io