Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods

Maria Ryskina*, Ella Rabinovich†, Taylor Berg-Kirkpatrick‡, David R. Mortensen* and Yulia Tsvetkov*
*{mryskina, dmortens, ytsvetko}@cs.cmu.edu †ella@cs.toronto.edu ‡tberg@eng.ucsd.edu

Summary

We analyze *neology*, the process by which new words emerge in a language, using large diachronic corpora of English. We compare language-internal and language-external factors by testing the following two hypotheses:

Supply: Neologisms are more likely to emerge in sparser areas of the semantic space
Demand: Neologisms are more likely to emerge in semantic neighborhoods of growing popularity

We find both factors to be predictive of word emergence although we find more support for the demand hypothesis.

Neologisms and control sets

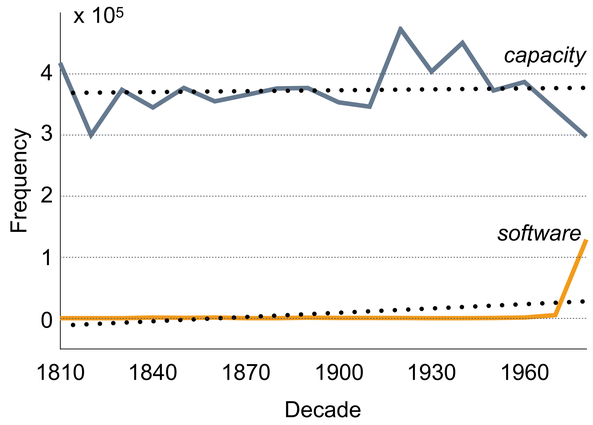
We select 1000 nouns that are 20x more frequent in the modern corpus (COCA) than in the historical one (COHA).

OED: ~58% of words (latest sense) emerged in the 20th century.

Neologism	$f_m \times 10^5$	$f_h \times 10^6$	*OED*	*Control*
voice-over	9.46	0.21	1966	*experience*
video	8.13	2.34	1981	*henry*
software	4.71	1.01	1958	*capacity*
gender	4.23	1.09	1984	*method*
e-mail	4.11	0	1979	*artist*
teaspoon	2.45	0.99	1791	*element*
infrastructure	1.66	0.33	1927	—
feedback	1.61	0.57	1943	*academy*
lifestyle	1.52	0.38	1929	*alliance*
…	…	…	…	…

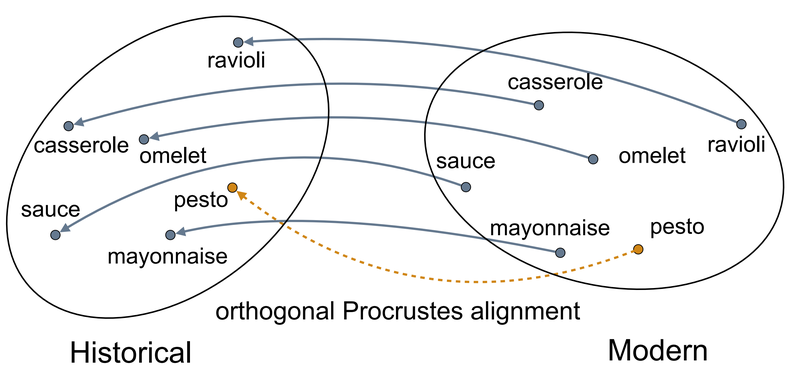
We pair each neologism with a control word, controlling for word length and frequency.

We use a *stable* and a *relaxed* control sets, i.e. with and without the stability constraint (absence of a monotonic frequency change pattern of a control word).



Methodology

1. Align embedding spaces by rotating and then project the neologisms into the historical embedding space:



2. Compare characteristics of the embedding space neighborhoods of the neologisms and the control words:

- Density: number of words in a neighborhood

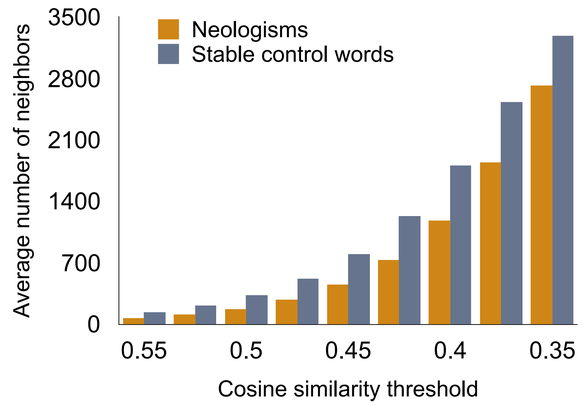
$$d(w,\tau) = |\{u : \text{cosine}(v_w, v_u) \geq \tau\}| \quad \Leftarrow \quad \text{supply}$$

- Frequency growth rate: average Spearman correlation between decades and frequency by decade in COHA

$$r(w,\tau) = \frac{1}{d(w,\tau)} \times \sum_{u:\text{cosine}(v_w,v_u)\geq\tau} r_s\left(\{1:18\}, f_{(1:18)}(u)\right) \quad \Leftarrow \text{demand}$$

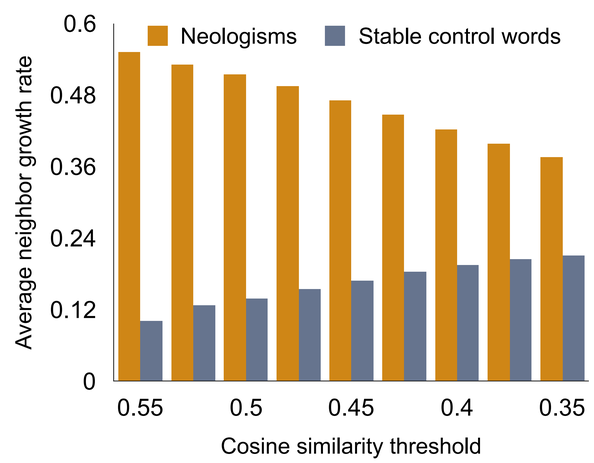
Results and analysis

GLM: $y(w) \sim \sigma\left(\beta_0^{(\tau)} + \beta_d^{(\tau)} \cdot \underbrace{d(w,\tau)}_{\text{density}} + \beta_r^{(\tau)} \cdot \underbrace{r(w,\tau)}_{\text{frequency growth}}\right)$



Supply: density is predictive with negative weight

Significant at $p < 0.01$ for larger neighborhood sizes (stable control set) or for all neighborhoods (relaxed control set).



Demand: frequency growth rate is predictive with positive weight

Significant at $p < 0.01$ for all neighborhood sizes and control sets.

Qualitative examples of nearest historical neighbors:

Neologism	*Nearest neighbors*	
email	*telegram*	*letter*
pager	*beeper*	*phone*
blogger	*journalist*	*columnist*
spokeswoman	*spokesman*	*director*
sushi	*caviar*	*risotto*
e-book	*paperback*	*hardcover*
hip-hop	*jazz*	*rock-n-roll*
daycare	*day-care*	*childcare*
vibe	*ambience*	*ambiance*
chemo	*chemotherapy*	*dialysis*