Supporting Information for

How Cognitive Selection Affects Language Change

Ying Li; Breithaupt Fritz, Cynthia S. Q. Siew, Thomas Hills, Ziyong Lin, Yanyan Chen and Ralph Hertwig

Corresponding Author Ying Li and Breithaupt Fritz.
E-mail: liying@psych.ac.cn or fbreitha@indiana.edu

This PDF file includes:

Supporting Information Text

Section 1: Multicollinearity check for Study 1 and Study 2

We first examined multicollinearity for all models. Independent variables in multiple regression models are supposed to be independent from each other so that researchers can single out the effect of one variable when others are controlled. Multicollinearity (i.e., high intercorrelations among two or more independent variables) makes it impossible and often leads to unreliable estimates of regression coefficients. Multicollinearity can be assessed by the variance inflation factor (VIF), which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model. We computed VIF for each independent variable. The smallest possible VIF value is 1, suggesting complete absence of multicollinearity. As a rule of thumb, a VIF value that exceeds 5 indicates a problematic amount of collinearity. We found that all independent variables had a VIF value smaller than 2 (SI Table S1).

Section 2:Preservation rate for words of different grammatical categories (Study 1)

See SI Figure S1.

Section 3: Relationship between increase of word frequency and six psycholinguistic properties

Comparing with COHA and Google English Fiction, Google Ngram All English shows divergent results from for arousal, concreteness and word length (SI Figure S2). As we elaborated in the SI section 8, we expected the difference because the Google Ngram All English corpus are overrepresented with scientific writing in recent decades, which arguably favours abstract words, neutral words, low arousal words and longer words. In contrast, COHA and Google English are free from such bias because of the fixed genre composition in each decade. Our result highlights that caution must be taken when making inferences on language and cultural change from historical corpora: 1) results should be interpreted in relation to the characteristics of the chosen corpus; 2) results should be cross-validated using two or more different corpora if possible.

Section 4: The result without screening out words that changed their meanings (corresponding to Figure 1 in the main text)

See SI Figure S3.

Section 5: Serial Reproduction Experiment

Figure S4a describes the procedure of the serial reproduction experiment. To prepare the original stories, Breithaupt et al. created three basic stories for each of five emotion categories (joy, sadness, disgust, embarrassment, risk). Next, for each of the above 15 basic stories, 5-8 variations were created by changing the content of the last sentence so that these variations differ in the intensity of its corresponding emotion*. In total there are 97 initial stories. Each participant read and retold three story variations after being instructed to *tell the story to another person in your own words*. Participants were not instructed to pay attention to any specific aspects of the stories. The three stories that each participant received are from different emotion categories and from within the same iteration (e.g. only second generations of retellings). In making use of the existing data, we removed participants who indicated that their first language is not English and who wrote in languages other than English. Following these procedures, the 97 original stories were retold for a total of 2,695 times in the first iteration, 6,474 times in the second iteration and 7,428 times in the third iteration. The stories get shorter as they propagate down the diffusion chain: from an average of 86 words in original stories to only 24 words in iteration 3 (Figure S4b For more details on the corpus see Breithaupt, Li, and Kruschke (1).

Section 6: Historical Corpora

The major difference among these three corpora is their genre composition. COHA (2) was carefully curated to be genre-balanced across decades, with major genres including TV/movies (subtitles), fiction, magazine, newspaper, and non-fiction. The corpus is also balanced across decades for sub-genres and domains (e.g., sub-genres for fiction are prose, poetry, drama, etc.). Fiction and TV/movies are the biggest genres in COHA, together accounting for 54-57% of the total in each decade. Google Ngram English Fiction includes only fiction books. The advantage of COHA and Google Ngram English Fiction is that they have fixed genre compositions across time. Hence, it alleviates the concerns that the historical patterns we observe (e.g. change of word frequency) are merely an artifact of changing genre compositions. Moreover, both corpora are large, with COHA containing around 475 million words from 1820 to 2000 and Google Ngram English Fiction containing around 40 billion words.

The third corpus, the Google Ngram All English corpus, is made up of around 155 billion words. It represents around 6% of all books published in English over the last several hundred years. On the one hand, this corpus has proved fruitful in capturing cultural shift such as evolution of grammar, adoption of technology (3) and national wellbeing (4); on the other hand, it has been criticized for containing corpus artifacts due to its shifting sampling paradigm. For example, Pechenick, Danforth,

*For example, in a story created under the emotion category "disgust", the authors manipulated the last sentence so that the story was perceived with slight disgust (the protagonist *eat leafy greens with fruit*) to strong disgust (the protagonist eat *soup that contains the whole fruit bat*). In a pretest, the researchers validated the effectiveness of the manipulation of emotion intensity by asking other participants to read the stories and evaluate the degree of intensity of its corresponding emotion and other measures.

and Dodds (5) found evidence for an explosion of scientific writings since around 1930s in Google Ngram All English, but not in Google Ngram English Fiction. This surge of academic writings offers a natural control for other corpora as we note below.

In Figure S5 we present additional evidence for an increase in scientific writing. Adding a suffix to the end of a word is a common strategy for word nominalization. It makes a word longer and is often perceived to be more professional (6). Figure 3 presents a summation of frequency for words that end with each of three common suffixes (-ism, -tion, -ity) and words longer than 15 letters. Unsurprisingly, Google Ngram All English contains much more nominalized words and longer words than COHA and Google Ngram English Fiction corpus, but the difference only becomes larger in recent decades. This further suggests that the Google Ngram All English corpus reflects an increase in scientific writing and should therefore show patterns of age of acquisition, concreteness, and emotional valence that are likely to be different from those found in our other two corpora. Thus, we provide the Google Ngram All English corpus for comparison in the Supplementary Materials (Figure S2 bottom). Here we focus on COHA and the Google Fiction corpus, which are not subject to this known sampling bias and therefore may better reflect the language use of its time period (SI Figure S5).

Section 7: Identifying Patterns of Word Frequency Change Using Principal Component Analysis

We are aware that historical word frequency may not always follow a monotonic trend. If there is a large proportion of words with non-monotonic frequency change, representing change of word frequency as the difference between two endpoints could be problematic. To find out the proportion of words with stable historical frequency, monotonic frequency change and non-monotonic frequency change, we used principal component analysis (PCA) on relative frequencies[†] between the years 1800 to 2000. PCA captures patterns of change without relying on prior assumptions. We grouped words that score at different ranges of each principal component (e.g., top 10%, 10%-20%, etc), and plotted the averaged frequency for each group (Figure S6). The first principal component (PC1) corresponds to the absolute frequency. Words that score highly on the opposite ends of PC1 are high frequency words like *all, one, on*, and low frequency words like *reformat, roadbed, histamine*. The second principal component (PC2) captures monotonic change: words with monotonically increasing frequency on one end of the PC2 and words with monotonically decreasing frequency on the other. The third principal component (PC3) represents an asymmetric curve, with U-curve on one end and inverted U-curve on the opposite end. In the Google English Fiction corpus, PC1 explains the most variance (48%), suggesting that rank order of word frequency is largely preserved for most words. PC2 explains around 19% of the variance. In contrast, PC3 only explains around 5% of the variance. Given the large variance explained by PC1 and PC2, our analysis focusing on word frequency difference between the years 1800 and 2000 captures the dominant pattern of word frequency change.

Section 8: Quantifying Semantic Stability

We quantified semantic stability following the procedure described in (7). The description from the original article is shown below.

Taking a Firthian approach (8), we assumed that the meaning of a word can be reliably inferred from the linguistic contexts in which the word has been used in. Therefore, the semantic shift of a word between two time points in history can be captured by comparing the extent to which its context has changed. This approach captures shift in two kinds of meanings: denotation (meaning that can be looked up in dictionary) and connotation (associations evoked by words in the mind of readers). For example, although the denotative meaning of woman (adult female human being) remained unchanged over the past 200 years, its connotation has become increasingly associated with gender equality movement. We obtained diachronic word embeddings trained on the Google Ngram Corpus from (9). These word embeddings were trained using SVD, which were constructed based on the following steps. First, a co-occurrence matrix was constructed to record the number of times any two words co-occurred within fixed-size sliding windows of text. Second, vectors containing the number of times a given word co-occurred with all other words were directly obtained from the co-occurrence matrix described above. Third, they computed the positive pointwise mutual information (PPMI) for each pair of words and then constructed a PPMI matrix with entries given by

$$\mathrm{PPMI}(v_i, v_j) = \max(0, \log(\frac{P(v_i, v_j)}{P(v_i) \times P(v_j)})) \quad [1]$$

where v_i, v_j represents a pair of words from the corpus. $p(v)$ corresponds to the empirical probabilities of word co-occurrences within a fixed-size sliding window of original text. As compared to co-occurrence counts, PPMI penalizes high-frequency words (i.e., *of, the, and*) that were used in a wide range of contexts, and favors word pairs that frequently appeared together but not with others (i.e., *Hong* and *Kong*). Forcing PPMI values to be above zero ensures that they remain finite and this has been shown to improve results (10). Finally, dimensionality of word embeddings was reduced to 300 using singular value decomposition (SVD). This dimensionality reduction acts as a form of regularization and allows us to compare word similarities by computing the cosine similarity of word embeddings.

With diachronic word embeddings, the semantic stability (i.e., the inverse of the rate of semantic change) of a given word can be quantified as

$$Semantic\ stability^{T1,T2}(w_i) = \mathrm{cos\_dist}(w_i^{(T1)}, w_i^{(T2)}) \quad [2]$$

[†] For each word, frequencies were divided by the historical maximum so that frequencies range from 0 to 1.

where $w_i^{(T)}$ refers to the word embedding of word w_i in year T. The historical embedding is aligned to its modern embedding using orthogonal Procrustes (11). Semantic similarity ranges from 0 to 1. For example, the semantic similarity of *happy* between year 1800 and 2000 is 0.73, much higher as compared to words that had undergone greater semantic change, such as *gay* (0.36), and *car* (0.41). Figure S7 (left) shows the distribution of semantic similarity between 1800 and 2000 for 50,000 English words trained on the Google Ngram Corpus. The negatively skewed distribution suggests that the majority of words were used in similar contexts at both time points. Figure S7 (right) shows the semantic stability of a few words as examples. Each line represents the semantic similarity between its historical meaning (across years 1800-1990) and its contemporary meaning (year 2000) of the corresponding word. The average semantic stability of the entire database is plotted in grey as a benchmark. Figure 1 (right) suggests that the word *happy* is relatively stable in its semantics over the past two centuries. In contrast, *gay* and *want* changed their meaning drastically (gay changed from happy to homosexuality; want changed from poverty/lack to the desire to do or possess)."

In the result we reported in Figure 1 (the main text), we remove all words (7.5% of the vocabulary) with semantic stability score (between 1800 and 2000) less than 0.4. We report the results with the unscreened full data in Figure S3.

Table S1. Variance Inflation Factor for each independent variables in Table 1

	Session	*AoA*	*Valence*	*Arousal*	*Concreteness*	*Emotionality*	*Length*	*Count*	*Frequency*	*Log Frequency 1800*
Study 1	1.0	1.7	1.2	1.2	1.4	1.3	1.4	1.0	1.7	N.A
Study 2	N.A	1.7	1.2	1.2	1.4	N.A	1.3	N.A	N.A	1.47

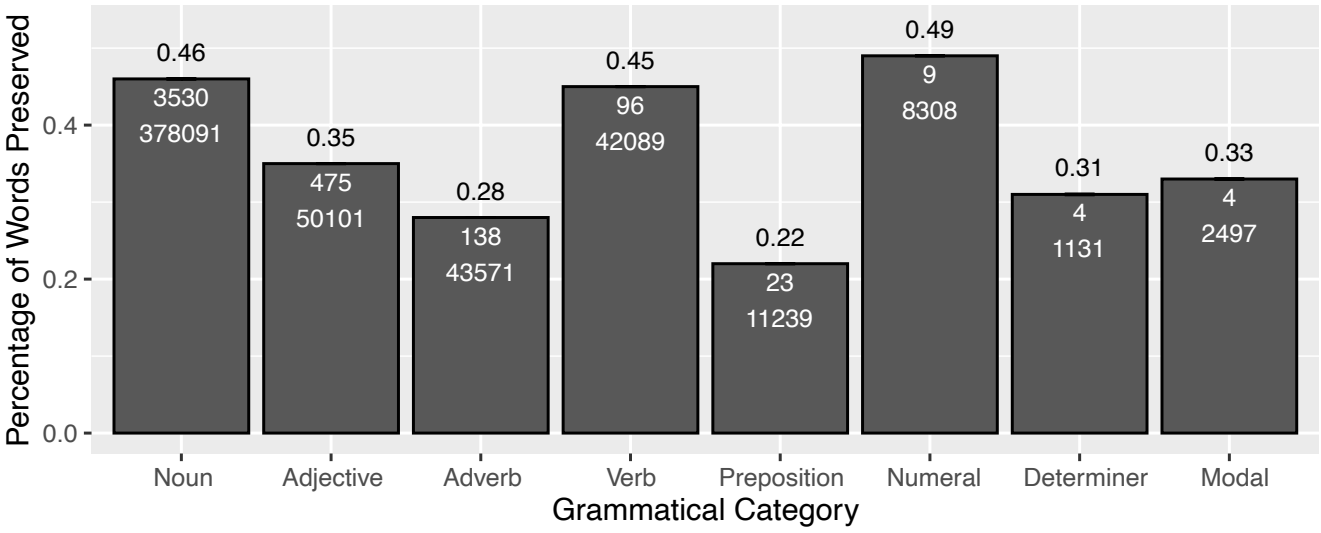
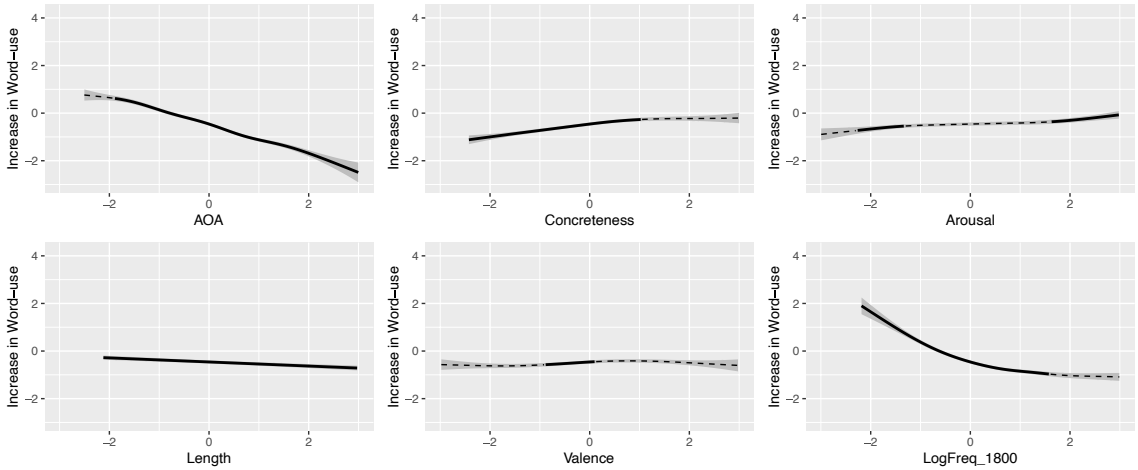
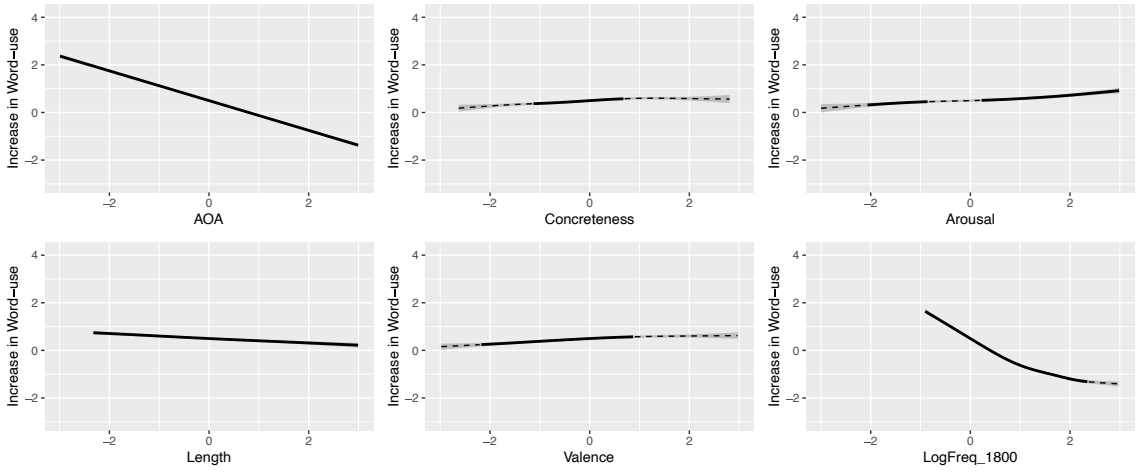


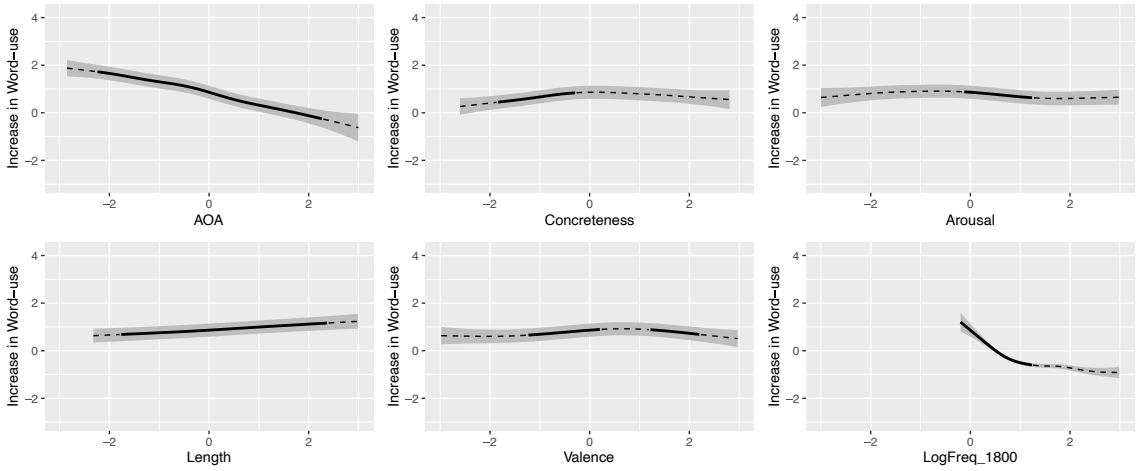
Fig. S1. The percentage of words preserved in story retelling task by grammatical category. The text label from top to bottom respectively represents the percentages of word preservation, number of word types and number of word tokens in the corresponding grammatical categories.



(a) COHA



(b) Google English Fiction



(c) Google All English

Fig. S2. Relationship between increase of log frequency between 1800 and 2000 (y-axis) and six psycholinguistic properties (x-axis). All independent variables are scaled and centred around 0. Lines are model estimates (dotted lines are estimates and solid lines indicate where the estimates are significantly increasing or decreasing). Shading indicates 95% confidence intervals around the estimates.

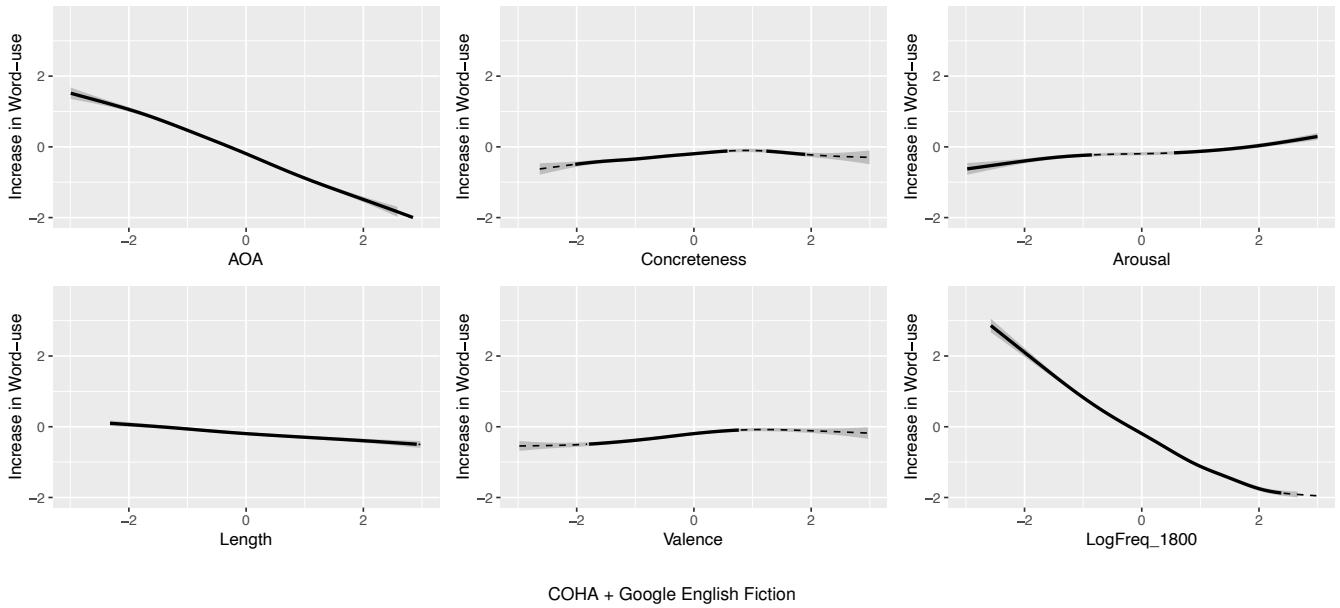


Fig. S3. Relationship between increase of log frequency between 1800 and 2000 (y-axis) and six psycholinguistic properties (x-axis). All independent variables are scaled and centred around 0. Lines are model estimates (dotted lines are estimates and solid lines indicate where the estimates are significantly increasing or decreasing). Shading indicates 95% confidence intervals around the estimates.

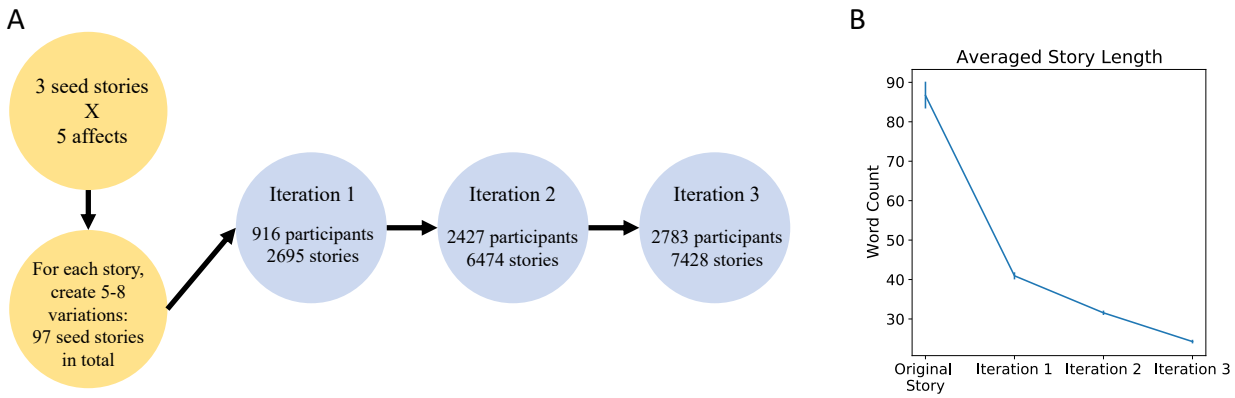


Fig. S4. (A) An illustration of the story retelling experiment (1). Different participants were recruited for each stage of the experiment. Narrators were asked to retell the story given to them in their own words. (B) The averaged story length in each iteration.

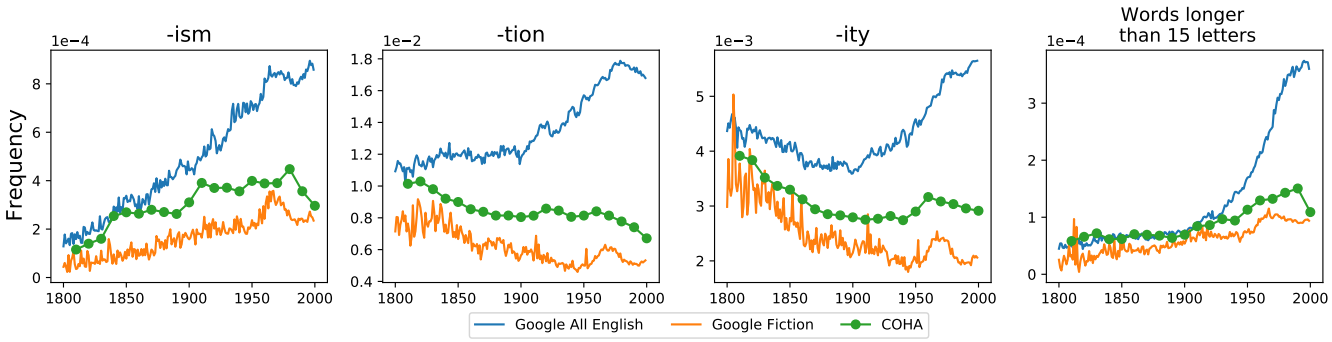


Fig. S5. Summation of historical frequency for words that end with each of the three common suffix for nominalization (-ism, -tion, -ity) and words longer than 15 letters.

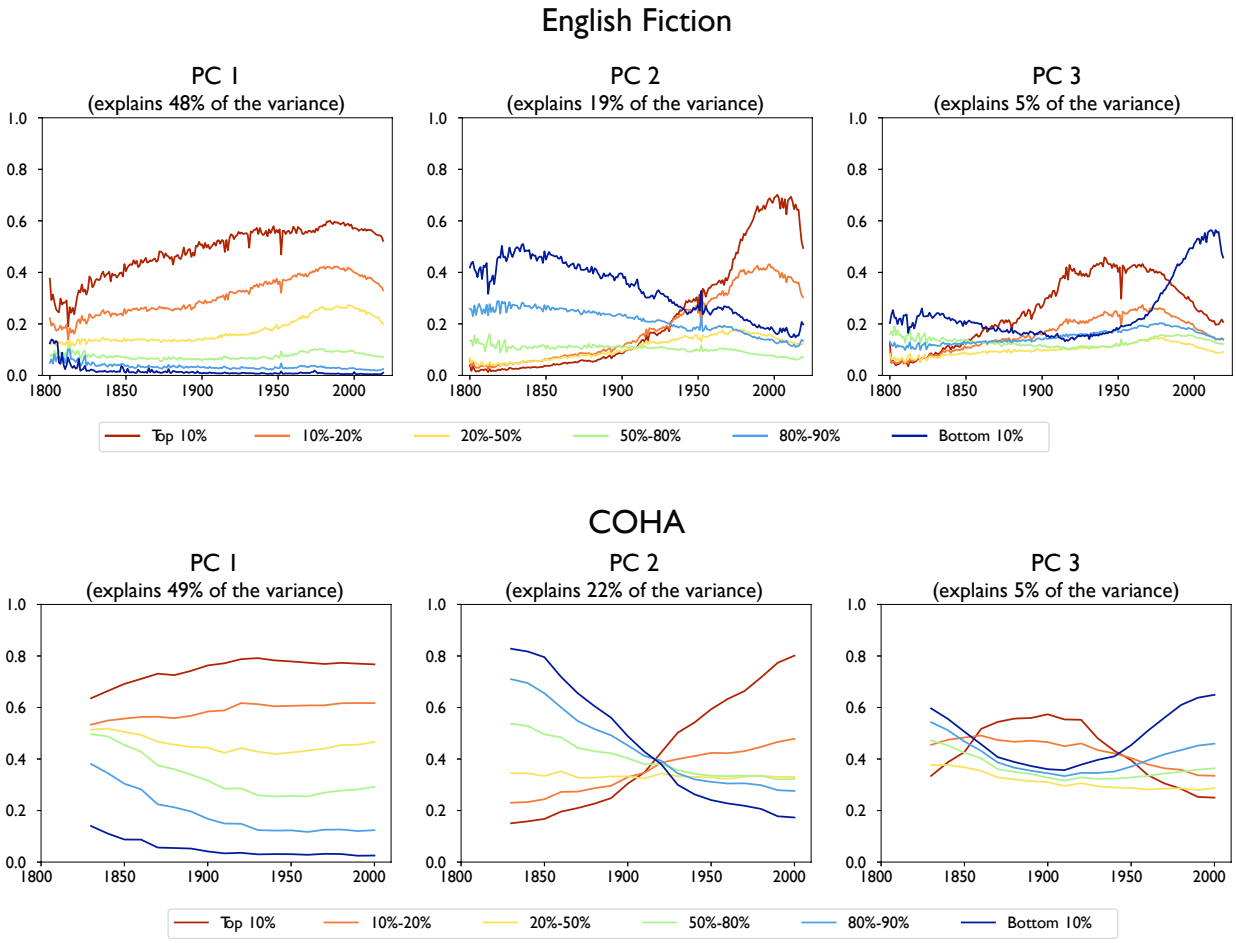


Fig. S6. Averaged frequency for words that score at different ranges of PC1, PC2 and PC3. The principal component analysis was conducted on relative word frequency between 1800 and 2000 from the Google English Fiction corpus and COHA. Respectively, PC1, PC2, and PC3 corresponds to word frequency, monotonic frequency change and non-monotonic frequency change.

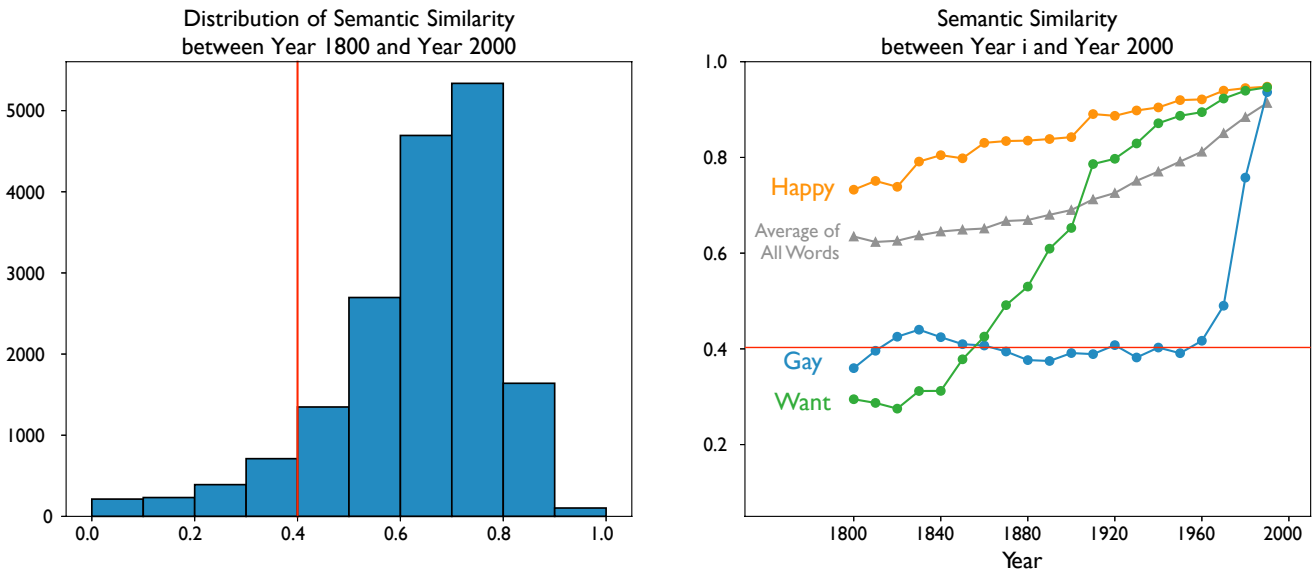


Fig. S7. Both figures were produced based on the Google Ngram Corpus. Left: Distribution of semantic similarities between 1800 and 2000. Right: Semantic stability of selected words. Each line represents the semantic similarity between the historical meaning (across years 1800-1990) and the contemporary meaning (year 2000) of the corresponding word. The grey line represents the average semantic similarity across all words in the Macroscope database (9). The red lines represent the threshold we used to screen out words that substantially changed their meaning.

References

1. F Breithaupt, B Li, JK Kruschke, Serial reproduction of narratives preserves emotional appraisals. *Cogn. emotion* **36**, 581–601 (2022).
2. M Davies, *The corpus of historical American English: COHA*. (BYE, Brigham Young University), (2010).
3. JB Michel, et al., Quantitative analysis of culture using millions of digitized books. *SCIENCE* **331**, 176–182 (2011).
4. TT Hills, E Proto, D Sgroi, CI Seresinhe, Historical analysis of national subjective wellbeing using millions of digitized books. *NATURE HUMAN BEHAVIOUR* **3**, 1271–1275 (2019).
5. EA Pechenick, CM Danforth, PS Dodds, Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one* **10**, e0137041 (2015).
6. H Sword, Zombie nouns. *The New York Times* p. 36 (2012, July 23).
7. Y Li, CS Siew, Diachronic semantic change in language is constrained by how people use and learn language. *Mem. cognition* **50**, 1284–1298 (2022).
8. JR Firth, *Papers in linguistics 1934–1951*. (Oxford University Press), (1957).
9. Y Li, T Engelthaler, CS Siew, TT Hills, The macroscope: A tool for examining the historical structure of language. *Behav. research methods* **51**, 1864–1877 (2019).
10. JA Bullinaria, JP Levy, Extracting semantic representations from word co-occurrence statistics: A computational study. *Behav. research methods* **39**, 510–526 (2007).
11. PH Schönemann, A generalized solution of the orthogonal procrustes problem. *PloS one* **31**, 1–10 (1966).