SEMANTIC BLEACHING NOT OBSERVED IN SYNCHRONIC TEST

Dillon Niederhut*[1]

*Corresponding Author: dillon.niederhut@gmail.com
[1]Enthought Inc., Austin, United States of America

It is well understood that the semantic content of words changes over time, but precisely how and why this happens is still unknown. Here, we test for synchronic evidence of semantic bleaching in a corpus of English language collected in 2016. We find no evidence of long-term reduction in the semantic value of words, although this may not be true when considered over shorter periods of time.

1. Introduction

SEMANTIC BLEACHING is a well-described phenomenon where the specificity of a word decreases with use. To take one common example, the term *awesome* was once reserved for the Judeo-Christian deity, but is now used to describe everything from toast[1] to the Transformer movie franchise[2]. Another way to state this observation is that a word which first refers one thing (which is presumably the case for all words) can be generalized over time to also refer to other, related things. In the example given here, *awesome* has been extended to include all sorts of things which supposedly inspire feelings of religious devotion.

Semantic bleaching may be part of a more general process whereby words expand and shift their meanings in order to maximize some definition of communicative optimum. Piantodosi, Tily, and Gibson argue that the driving force is the production of speech, and that it is optimal to find new uses for phrases with short orthographic length in order to reduce the overall number of graphemes or syllables needed to convey any particular idea (Piantadosi, Tily, & Gibson, 2011). Recent work by Xu has also incorporated the cost of interpretation of words, focusing on reducing the ambiguity during the process of assigning labels to objects. In one semantic domain, the historical shift in word meanings over time approaches a Pareto frontier balancing the cost of production with the cost of interpretation for assigning existing words to new kinds of containers (Xu, Regier, & Malt, 2016).

[1]http://www.bonappetit.com/restaurants-travel/article/how-to-make-perfect-toast

[2]http://kotaku.com/leave-michael-bay-alone-transformers-is-awesome-1596887614

DOI:10.12775/3991-1.082

Here, we conduct a synchronic test for historical evidence of semantic bleaching in an English corpus collected in 2016. Under the assumption the rate of bleaching per word tends to be positive, i.e. that it outpaces the rate of fossilization, then it follows that in aggregate older words should have less specific meaning than newer words. This must be certainly be true in the narrow sense, at least for the coinage of new terms, but it remains to be seen whether this relationship holds over historic time.

2. Methods

The date of first written appearance of a large number of English words was acquired from Merriam Webster's recently published "Words by First Known Date" (Merriam Webster, 2018). This generated dates for approximately 10,000 terms, with a roughly even distribution over the last 100 years, and appearances of terms becoming less specific further than the 1800s (Fig. 1)

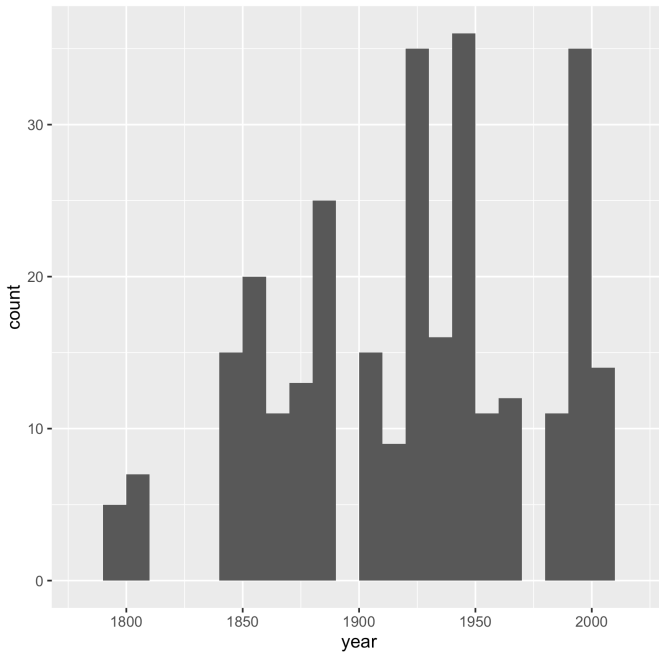

Figure 1. Distribution of estimates of the year of first appearance for 1,000 English words; each bin has a width of ten years

Term frequency was acquired from redicorpus dataset described in Niederhut

Niederhut (2017). Briefly, this corpus includes approximately 4.3 million communication events sampled from the AskReddit subgroup of reddit.com over 22 weeks in 2016. We were able to discover daily use frequencies and communication contexts for approximately 2500 of the terms from Merriam Webster after applying the Porter Stemmer (Porter, 1980). While redicorpus contains terms of all kinds (including function words), the matches with the Merriam Webster data are largely nouns and verbs like *unfriend*, *mantissa*, and *carotid sinus*.

Zipf statistics, a measure of semantic value, were calculated according to the formula from (Niederhut, 2016). The intuition is that words like *multicollinear* feel like they contain more information than words like *my*, because the former provides evidence of a rather specific context (a problematic linear model) whereas the latter does not. The belief that an utterance comes from a specific linguistic context creates an expectation that other context-appropriate terms like *variance inflation factor* will also be observed. The Zipf statistic quantifies the magnitude of this expectation by comparing the probability distribution of words conditioned on one context against the distribution of words across all contexts.

We were able to calculate Zipf statistics for about 300 of the Merriam Webster terms. This number is remarkably smaller due to the computational cost of calculating these statistics. Each test statistic takes roughly 90 minutes to process, largely due to the I/O cost of reading large amounts of text data. The sample thus represents roughly 435 CPU hours, or eighteen days, of work.

Table 1. Summary statistics for year of first appearance, frequency of use in 2016, and estimated Zipf statistic

	year	frequency	Zipf statistic
Min.	1799	9.3e-09	-0.794
1st Qu.	1879	2.8e-08	1.495
Median	1926	1.3e-07	2.534
Mean	1919	1.6e-07	7.354
3rd Qu.	1954	6.5e-07	4.653
Max.	2009	8.3e-04	233.577
N	10191	2594	290

A linear model was run regressing the Zipf statistics on the year of first appearance of each word. To help control for the effect of any outliers, the model was rerun using the robust linear modeling package for R (Wang et al., 2014). To test for evidence of semantic bleaching over short time scales, a third model was run including only data since the 1990s. To control for the possibility that the outcome might be determined by some peculiarity of the Zipf statistic, we also tested for a relationship between the age of a term and its daily moment[3].

[3]The daily moment of a term is it's average daily uses divided by the standard deviation in its daily

Data were collected in Python 3.5.4 on Ubuntu Server 16.0.4, and were analyzed with Revolution R Open[4] based on CRAN release v. 3.2.3, "Wooden Christmas Tree", (R Core Team, 2015). Tables were produced with xtable, and figures were produced with ggplot2 (Dahl, 2014; Wickham, 2009). Data and R files to reproduce this analysis along with its tables and figures are available at `https://github.com/deniederhut/semantic-bleaching-not-observed-in-synchronic-test`.

3. Results

Table 2. Model statistics from an OLS regressing Zipf statistic on year of appearance.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-25.42449	45.23668	-0.562	0.575
year	0.01706	0.02353	0.725	0.469
Residual standard error: 21.65 on 288 degrees of freedom				
Multiple R-squared: 0.001821, Adjusted: -0.001645				
F-statistic: 0.5255 on 1 and 288 DF, p-value: 0.4691				

We find no evidence that the Zipf statistic of a word is related to how long that word has been in use. The first linear model assigns a coefficient of 0.03 to the year term, which is not significantly different than zero at $p = 0.20$ (Table 2). The R^2 for the model, in both the corrected and uncorrected estimates, is less than 1%.

The robust model, which ignores roughly 40 of the very large Zipf values, produces similar results, with a coefficient for the year term that is less than 0.01, and not significantly different from zero with $p = 0.38$ (see Fig. 2). The R^2 for the model, in both the corrected and uncorrected estimates, is less than 1%.

Table 3. Model statistics from a robust model regressing Zipf statistics on years after 1990.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-154.23339	116.92727	-1.319	0.194
year	0.07844	0.05865	1.337	0.188
Robust residual standard error: 1.766				
Multiple R-squared: 0.04715, Adjusted R-squared: 0.02499				

The recent model, which includes only the last twenty years in the data set, offers some mild evidence toward a short-term effect of time on semantic value,

use, and can be interpreted as a measure of generality. Common words like *deny* have large moments (here, cerca 3.2), while uncommon words like *bantamweight* have small moments (here, c. 0.08)

[4]https://mran.revolutionanalytics.com/open/

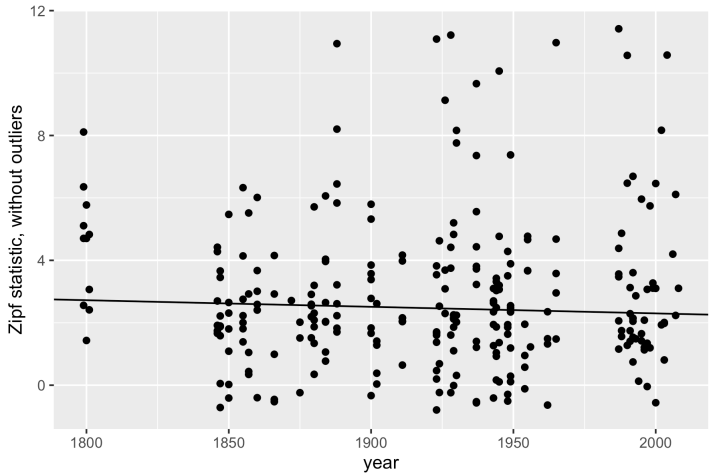

Figure 2. Relationship between Zipf statistics and year of appearance, with outliers (as determined by robust algorithm) removed, with superimposed trend line from robust model.

with a positive effect coefficient of 0.08 and $p < 0.20$ (Table 3). The R^2 for the model, in both the corrected and uncorrected estimates, is roughly 4%.

Table 4. Model statistics from an OLS regressing daily moment on year of appearance.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0008698	1.1729085	0.001	0.999
year	0.0002458	0.0006102	0.403	0.687
Residual standard error: 0.5612 on 288 degrees of freedom				
Multiple R-squared: 0.0005633, Adjusted: -0.002907				
F-statistic: 0.1623 on 1 and 288 DF, p-value: 0.6873				

The comparison against the daily moment produces the same null result as the overall and robust tests against the Zipf statistic, with no significant relationship between the year of first appearance of a term and the generality with which it is used in natural language production.

4. Discussion

It is interesting that we find no relationship between the semantic value of a word, as measured by the Zipf statistic, and how long that word has been in use. Prima facie, the oldest words in a language should have had more time to take on additional meanings, and to have had their specificity diluted through use in metaphor and other poetic devices. One interpretation of this finding is that bleaching takes

different trajectories for different words, and that these changes will only be visible in diachronic tests.

Another possibility is that semantic bleaching, where the specificity of a word decreases over time, is balanced by a force that removes less-used senses of the word in order to reduce ambiguity in its meaning. This may be a passive process, where some semantic interpretations fall out of use simply because their referents do, like the use of *wire* to describe information transfer over telegraph.

A final interpretation is that Merriam Webster has chosen particularly interesting words about which to publish the year of appearance online. Based on prior work, we would expect a selection of words randomly sampled from human usage to have words that appear much more frequently. The words in this sample have an average proportion of $2.2e-06$, and a median of $1.3e-07$ (see Table 1). The presence of words like *wiki* and *a tempo* may also explain the unusually high Zipf statistics observed for these data.

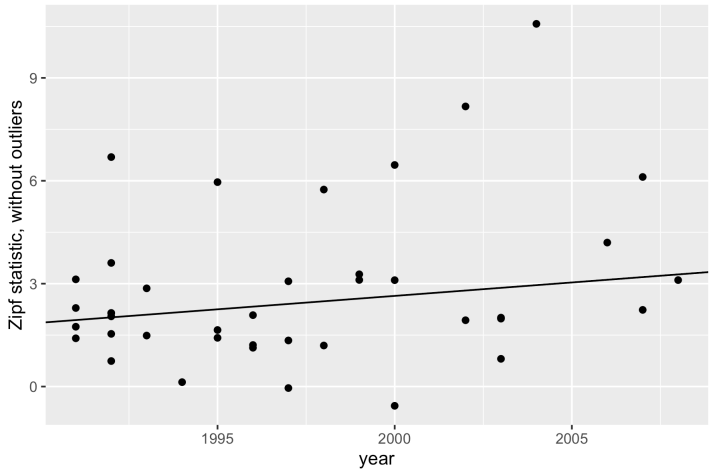

Figure 3. Relationship between Zipf statistics and year of appearance since the 1990s, with outliers (as determined by robust algorithm) removed, with superimposed trend line from robust model.

Given the findings from the third model, it remains possible that words bleach very quickly over a short duration, and then reach some kind of equilibrium in the population whose stable point is governed by other forces, likely including the size of the speech community (Fig. 3). However, the magnitude of the "short term bleaching" effect that we have measured is small enough that we do not feel comfortable arguing that it provides a plausible alternative to the definite null effect over longer periods of time without corroborating evidence of its existence.

References

Dahl, D. (2014). *xtable: Export tables to latex or html.* (R package version 1.7-4)

Merriam Webster. (2018). *Words by first known use date.* https://www.merriam-webster.com/words-by-first-known-date.

Niederhut, D. (2016). Quantifying the semantic value of words. In S. Roberts, C. Clusky, L. McCrohon, L. Barcelo-Coblijn, O. Feher, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 11th international conference.* Hackensack: World Scientific.

Niederhut, D. (2017). *Performance approaches to semantics in human language.* Unpublished doctoral dissertation, University of California at Berkeley.

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Science*, *108*, 3526-3529.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*, 130-137.

R Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria.

Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Marrona, R., Zivot, E., Rocke, D., Martin, D., Maechler, M., & Konis, K. (2014). *robust: Port of the s+ "robust" library.* (R package version 0.4-16)

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.* Springer New York.

Xu, Y., Regier, T., & Malt, B. (2016). Historical semantic chaining and efficient communication: the case of container names. *Cognitive Science*, *40*, 2081-2094.