Exploring the Google Books Corpus:  
An Information-Theoretic Approach to Linguistic Evolution  

Eitan Adam Pechenick, Christopher M. Danforth, Peter Sheridan Dodds  
Department of Mathematics and Statistics &  
The Computational Story Lab  
University of Vermont  
April 23, 2015  

Abstract:  

It is tempting to treat frequency trends from Google Books data sets as indicators for the true popularity of various words and phrases. Doing so allows us to draw novel conclusions about the evolution of public perception of a given topic. However, sampling published works by availability and ease of digitization leads to several important effects, which have typically been overlooked in previous studies. One of these is the ability of a single prolific author to noticeably insert new phrases into a language. A greater effect arises from scientific texts, which have become increasingly prolific in the last several decades and are heavily sampled in the corpus. The result is a surge of phrases typical to academic articles but less common in general, such as references to time in the form of citations. We highlight these dynamics by examining and comparing major contributions to the statistical divergence of English data sets between decades in the period 1800–2000. We find that only the English Fiction data set from the second version of the corpus is not heavily affected by professional texts, in clear contrast to the first version of the fiction data set and both unfiltered English data sets. We critique a method used by authors of an earlier work to determine the birth and death rates of words in a given linguistic data set. While intriguing, the method in question produces an artificial surge in the death rate at the end of the observed period of time. To avoid boundary effects in our own analysis of asymmetries in language dynamics, we observe the volume of word flux across various relative frequency thresholds for the second English Fiction data set. Finally, we contrast properties of exogenous signals, such as wars, with those of endogenous signals.