The Evolution of Written Language:
An analysis of the Google Books corpus

Eitan Pechenick
Dr. Peter Dodds

Department of Mathematics and Statistics &
The Computational Story Lab
University of Vermont

April 23, 2013

Abstract:

A wealth of data is available for the study of the volume and frequencies of words and phrases over time. In particular, the Google Books corpus represents a compilation of billions of word and phrase occurences in written works spanning five centuries and several languages. The sheer volume of these records, particularly over the last two hundred years, presents many opportunities for the analysis of the patterns in the evolution of written languages. This project examines the magnitude of the differences (the statisitical divergence) between temporal frames in a language—English, primarily—as well as the dominant contributions to these changes. By examining the effects on both a language as a whole and on specific samples of words and phrases, and by investigating other published analyses of the data, one can bring trends into focus and further qualify the nature of these shifts. Such trends include an overall decline in the yearly rate of divergence. Moreover, specific effects are apparent, including those from major conflicts, as well as the rapid rise of technology and the vocabulary it has introduced.