2015

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

Eitan Pechenick
University of Vermont

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

A Dissertation Presented

by

Eitan Adam Pechenick

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fullfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Mathematical Sciences

May, 2015

Defense Date: March 27, 2015
Dissertation Examination Committee:

Peter Dodds, Ph.D., Advisor
Jacques Bailly, Ph.D., Chairperson
Christopher Danforth, Ph.D.
Richard Foote, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

Abstract

The Google Books corpus contains millions of books in a variety of languages. Due to this incredible volume and its free availability, it is a treasure trove that has inspired a plethora of linguistic research.

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 appears to produce an artificial surge in the death rate at the end of the observed period of time. In order 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 (in both directions) for the second English Fiction data set. We then use the contributions of the words crossing these thresholds to the Jensen-Shannon divergence between consecutive decades to resolve major factors driving the flux.

Having established careful information-theoretic techniques to resolve important features in the evolution of the data set, we validate and refine our methods by analyzing the effects of major exogenous factors, specifically wars. This approach leads to a uniquely comprehensive set of methods for harnessing the Google Books corpus and exploring socio-cultural and linguistic evolution.

Citations

Material from this dissertation has been submitted for publication in *PLOS ONE* on 01/05/2016 in the following form:

E. A. Pechenick, C. M. Danforth, and P. S. Dodds (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE*.

AND

Material from this dissertation has been submitted to the *arXiv* on 03/11/2015 in the following form:

E. A. Pechenick, C. M. Danforth, and P. S. Dodds (2015). Is language evolution grinding to a halt?: Exploring the life and death of words in English fiction. *arXiv preprint*.

Dedicated to

in memory of

Isaac Asimov (1920-1992)

Acknowledgements

I take this opportunity to thank my friends and family—especially my parents, who have supported me always, and my brother, Dov, who helped me get focused when I needed it most. I would also like to thank my advisor, Peter Dodds, for his guidance, support, and patience; Christopher Danforth for acting as a second advisor for all intents and purposes; Richard Foote, for supporting my endeavors since I was a teenager and for his uncanny ability to turn everything into algebra (and vice-versa); Jacques Bailly, for agreeing to chair my defense committee on short notice; Andrea Elledge, for handling the logistics with unparalleled efficiency; David Van Horn, Josh Auerbach, Andy Reagan, Nick Allgaier, and Jake Williams who produced this document's template and Peter Dodds for Perl/LaTeX wizardry; Nick Allgaier, Cathy Bliss, and Jake Williams for all of their thoughts and guidance as graduate students; my teachers and professors, for their insights; my peers from over the years, for sharing in this enterprise; Upton Sinclair for creating Lanny Budd, who turns up astonishingly often herein; the dogs in the office, Finnigan and Banjo, just for being fuzzy; my brother's dog, Appa, for being incredibly fluffy; and my niece, Evan, for being just plain adorable.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

In this chapter, we introduce the Google Books corpus, which is both incredibly large and freely available to the public. We then discuss previous works that examine the properties of this linguistic corpus. Although some analyses of the corpus focus on the frequency time series of prechosen words and from these draw broad conclusions, several related works explore dynamic properties on the scale of an entire language. However, even these sometimes draw conclusions without first exhaustively considering the nature of the data sets examined, which means that additional context is required to properly interpret the results of these studies. Moreover, many of these analyses were performed using the original version of the Google Books corpus, which was published in 2009, instead of the more recent 2012 incarnation, which also creates opportunities for review of these previous works, as well as the basis for exploring the evolution

of the Google Books corpus between versions. Finally, we introduce tools from information theory that allow us to explore the dynamics of these data sets and provide the context necessary to draw strong conclusions about socio-linguistic evolution on the basis of this corpus.

1.1 Introduction

J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

The Google Books corpus is freely available and contains an incredible volume of words and phrases. This makes it a natural subject for research in computational linguistics. The first version of the data set, published in 2009 [1], distills a collection of 15 million digitized books, mostly provided by university libraries, into a selection of 5 million books determined to be suitable with regard to optical character recognition accuracy, availability and accuracy of metadata, and and other factors. These 5 million books contain over half a trillion words. Of these, 361 billion are in English. The 2009 version also incorporates texts in Spanish, French, German, Russian, Chinese, and Hebrew. As well as producing this corpus and making it available to the public, the authors of [1] also performed various analyses, as we will include in Section 1.2, on their corpus to advance the field of culturomics.

The second version, published in 2012 [2], incorporates 8 million books. Although this version also includes syntantic annotations of many words as well as books in Italian, the importance of the second version to our work mainly arises from the fact that it contains half a trillion words in English alone. It should also be noted that the English Fiction data set in the 2012 version of the corpus is distinctly improved from the 2009 version of English Fiction, as we will demonstrate in Ch. 2.

1.2 Related Works

The simplest analyses involving Google Books track the relative frequencies of a specific set of words or phrases. Examples include phrases surrounding individuality [3], gender [4], urbanization [5], and time [1, 6]. However, such analyses focus on a relatively narrow set of words and phrases—anywhere from two [5] to twenty [3] or more at a time in the examples listed.

Many researchers have carried out analyses on entire data sets in the Google Books corpus. These include analyses of Zipf's and Heaps' laws as applied to Google Books data sets [7], rates of verb regularization [1], rates of word birth and death [6], and the tendency for several languages to change less and less over time [8].

However, these studies appear to ignore important details regarding the Google Books corpus. As we will demonstrate in Ch. 2, the sampling of texts in several English data sets in the corpus is biased toward the inclusion of scientific journals. This effect is especially prevalent during recent decades, which is of particular importance to all analyses concerned with recent social change.

A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley, "Statistical laws governing fluctuations in word use from word birth to word death," *Scientific reports*, vol. 2, 2012.

Furthermore, in Ch. 3, we critique the method employed in [6] to characterize the rates of word "birth" and "death" in various languages over time. The method in question is not without merit, and it motivates our explorations in that chapter. However, it requires an abundance of parameters, ties the birth and death of a word to its own relative frequency—so that each word is judged by a different threshold—and is subject to a major boundary effect. Namely, it always produces an increase in the death rate of words in the decades leading to the present. It must also be noted that one of the data sets analyzed in this paper is the 2009 version of English Fiction, which is not immune from the afore-mentioned sampling bias.

1.3 Information Theory

C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 1948.

The foundation of information theory dates to a seminal 1948 paper [9] in which the content of a physically transmitted signal is abstracted into a mathematical concept of information. Specifically, if the signal in question has a probability p of being observed, then that signal contains $-\log_2 p$ bits of information. This value decreases as p increases, and in the extreme cases of $p = 0$ and $p = 1$, a signal has infinite and zero information, respectively. Briefly stated: the rarest signals contain

the most information.

Beyond the mathematical notion of information, this paper also introduces Shannon entropy in analogy with physical entropy. This is defined as the average amount of information received (or transmitted) over a channel per signal, namely

$$H(P) = \sum_{i=1}^{N} -p_i \log p_i, \tag{1.1}$$

where P is the distribution of the probabilities of observing each signal, N is the number of possible signals, and p_i is the probability of receiving the i^{th} signal. Since the signal set may be of any variety—anything from characters in an alphabet to words in a language to species of frog in the Amazon rainforest—Shannon entropy is a powerful and broadly applicable tool.

One consequence of Shannon entropy and its relationship to physical channels is that given a set of signals that may be transmitted, one most efficiently uses the available bandwidth by encoding the signal set so that the length of each encoded signal contains the information content of that signal. Note that this also applies to efficiently recording and storing observed signals, namely compression. While one must in practice round to an integer number of bits, this theoretical construct nonetheless provides the framework necessary to measure differences between two distributions, P and Q, over the same signal set.

Kullback-Leibler divergence [10] between P and Q is defined as

$$D_{KL}(P||Q) = \sum_{i=1}^{N} p_i \log_2 \frac{p_i}{q_i}, \tag{1.2}$$

which we may rewrite as

$$D_{KL}(P||Q) = -\sum_{i=1}^{N} q_i \log_2 p_i - H(P). \tag{1.3}$$

The first term represents the average bandwidth use if the signals are distributed according to P but are encoded according to the distribution Q. Therefore, the Kullback-Leibler divergence can be interpreted as the number of bits wasted per signal. This waste analogy provides a tangible sense of the extent to which two distributions differ. However, Kullback-Leibler divergence has two properties that are not always desirable. First, it is not symmetric. Second, if any one signal occurs with nonzero probability in P but does not occur in Q, then the divergence between P and Q is infinite. This latter issue is of particular concern when dealing with distributions of words in a given linguistic data set as measured in two different time periods, since words may appear and disappear from a language—especially when optical character recognition errors are recorded as words.

Both of these issues are resolved if we instead use Jensen-Shannon divergence [11], which is defined as

$$D_{JS}(P||Q) = \frac{1}{2}\Big(D_{KL}(P||M) + D_{KL}(Q||M)\Big), \tag{1.4}$$

where $M = \frac{1}{2}(P + Q)$ is a mixed distribution. Unlike Kullback-Leibler divergence, Jensen-Shannon divergence is bounded. When P and Q are identical distributions, the Jensen-Shannon divergence is 0 bits. When P and Q share no overlapping signals, the

divegence is 1 bit. Furthermore, we may also express the Jensen-Shannon divergence as

$$D_{JS}(P||Q) = H(M) - \frac{1}{2}\Big(H(P) + H(Q)\Big), \tag{1.5}$$

from which it may be observed that a similar waste analogy holds as with Kullback-Leibler divergence. These properties and the ability to resolve the Jensen-Shannon divergence into contributions from each signal by way of Eq. 1.1 are at the core of our methodology in exploring the dynamics of the Google Books corpus. As such, these properties are reiterated and expounded upon in Chs. 2 and 3.

Chapter 2

Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution

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, such as time and gender. However, sampling published works by availability and ease of digitization leads to several important effects. One of these is the surprising 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. Here, 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. Our findings emphasize the need to fully characterize the dynamics of the Google Books corpus before using these data sets to draw broad conclusions about cultural and linguistic evolution.

2.1 Introduction

The Google Books data set is captivating both for its availability and its incredible size. The first version of the data set, published in 2009, incorporates over 5 million books. [1] These are, in turn, a subset selected for quality of optical character recognition and metadata—e.g., dates of publication—from 15 million digitized books, most of which were provided by university libraries. These 5 million books contain over half a trillion words, 361 billion of which are in English. Along with separate data sets for American English, British English, and English Fiction; the first version also includes Spanish, French, German, Russian, Chinese, and Hebrew data sets. The second version, published in 2012 [2], contains 8 million books with half a trillion words in English alone, and also includes books in Italian. The contents of the sampled books are split into n-grams, which are typically blocks of text

separated into n pieces by whitespace—e.g., "I" is a 1-gram, and "I am" is a 2-gram. (Phrases in the data sets are case-sensitive.) The resulting Google Books data sets, in which the n-gram lengths range from one to five, contain some exceptions to this rule of thumb. For example, in the 2009 version, "I am." is tokenized as the 3-gram, "I am ." The period is counted in the corpus as a 1-gram, as are other punctuation marks. Similarly, concatenations are often tokenized as 2-grams, so "I've been" becomes "I 've been"—also a 3-gram.

For ease of comparison with related work, we focus primarily on 1-grams from selected English data sets between the years 1800 and 2000. In this work, we will use the terms "word" and "1-gram" interchangeably for the sake of convenience. The total volume of (non-unique) English 1-grams grows exponentially between these years, as shown in Fig. 2.1, except during major conflicts—e.g., the American Civil War and both World Wars—when the total volume dips substantially. We also observe a slight increase in volume between the first and second version of the unfiltered English data set. Between the two English Fiction data sets, however, the total volume actually decreases considerably. This effect indicates insufficient filtering was used in producing the first version and immediately suggests the initial English Fiction data set may not be appropriate for any kind of analysis.

The simplest possible analysis involving any Google Books data set is to track the relative frequencies of a specific set of words or phrases. Examples of such analyses involve words or phrases surrounding individuality [3], gender [4], urbanization [5], and time [1, 6], all of which are of profound interest. However, the strength of all conclusions drawn from these must take into account both the number of words and

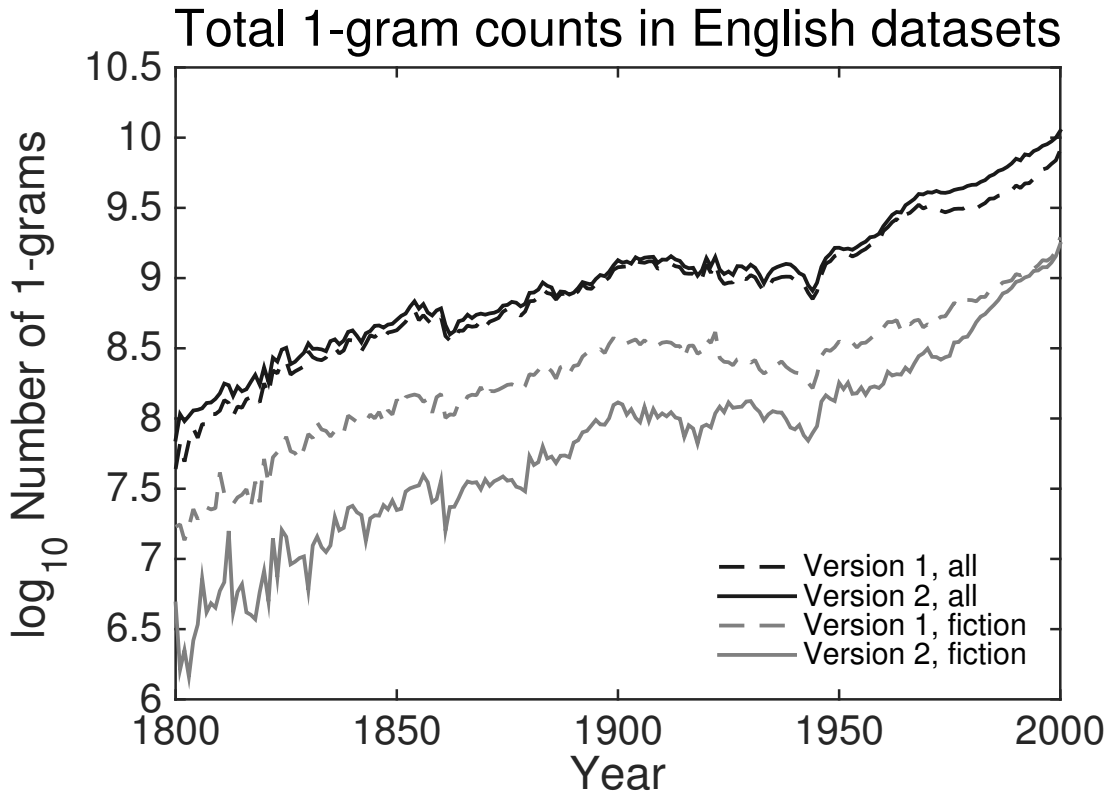

Figure 2.1: The logarithms of the total 1-gram counts for the Google Books English and English Fiction data sets are represented by the dark and light gray curves, respectively. The dashed and solid curves denote the 2009 and 2012 versions of the data sets, respectively. In all four examples, an exponential increase in volume is apparent over time with notable exceptions during wartime when the total volume decreases. (This effect is clearest during the American Civil War and both World Wars.) However, while the total volume for English increases between versions, the volume for fiction decreases drastically, suggesting a more rigorous filtering process.

phrases in question (anywhere from two [5] to twenty [3] or more at a time) and the sampling methods used to build the Google Books corpus.

Many researchers have carried out broad analyses of the Google Books corpus, examining properties and dynamics of entire languages. These include analyses of Zipf's and Heaps' laws as applied to the corpus [7], the rates of verb regularization [1], rates of word introduction and obsolescence and durations of cultural memory [6], as well as an observed decrease in the need for new words in several languages [8]. However, these studies also appear to take for granted that the data sets sample in a consistent manner from works spanning the last two centuries.

As we will demonstrate, an assumption of unbiased sampling of books is not reasonable during the last century and especially during recent decades, which is of particular importance to all analyses concerned with recent social change. Since parsing in the data sets is case-senstive, we can give a suggestive illustration of this observation in Fig. 2.2, which displays the relative frequencies of "figure" versus "Figure" in both versions of the corpus and for both English and English Fiction. In both versions of the English data set, the capitalized version, "Figure," surpasses its lowercase counterpart during the 1960s. Since the majority of books in the corpus originated in university libraries [1], a major effect of scientific texts on the dynamics of the data set is quite plausible. This trend is also apparent—albeit delayed—in the first version of the English Fiction data set, which again suggests insufficient filtering during the compilation process for that version.

Analysis of the emotional content of books suggests a lag of roughly a decade between exogenous events and their effects in literature, which may represent the time between the formative experiences of an author and the first publications of said

author [9]. Whatever the reason for the lag, however, it complicates the use of the Google Books data sets directly as snapshots of cultural identity. Similarly, authors are not represented equally in any given data set but are instead roughly sampled by prolificity. This leaves room for individual authors to have noteworthy effects on the dynamics of the data sets, as we will demonstrate in Section 2.3.

Lastly, due to copyright laws, the public data sets do not include metadata (see supporting online material [1]), and the data are truncated to avoid inference of authorship, which severely limits any analysis of censorship [1, 10] in the corpus. Under these conditions, we will show that much caution must be used when employing these data sets—with a possible exception of the second version of English Fiction—to draw cultural conclusions from the frequencies of words or phrases in the corpus.

We structure the remainder of the paper as follows. In Section 2.2, we describe how to use Jensen-Shannon divergence to highlight the dynamics over time of both versions of the English and English Fiction data sets, with particular attention given to key contributing words. In Section 2.3, we display examples of these highlights and reflect on the findings of this paper and means by which certain biases in the corpus may be overcome in future works. We offer concluding remarks in Section 2.4, summarizing the implications of our findings.

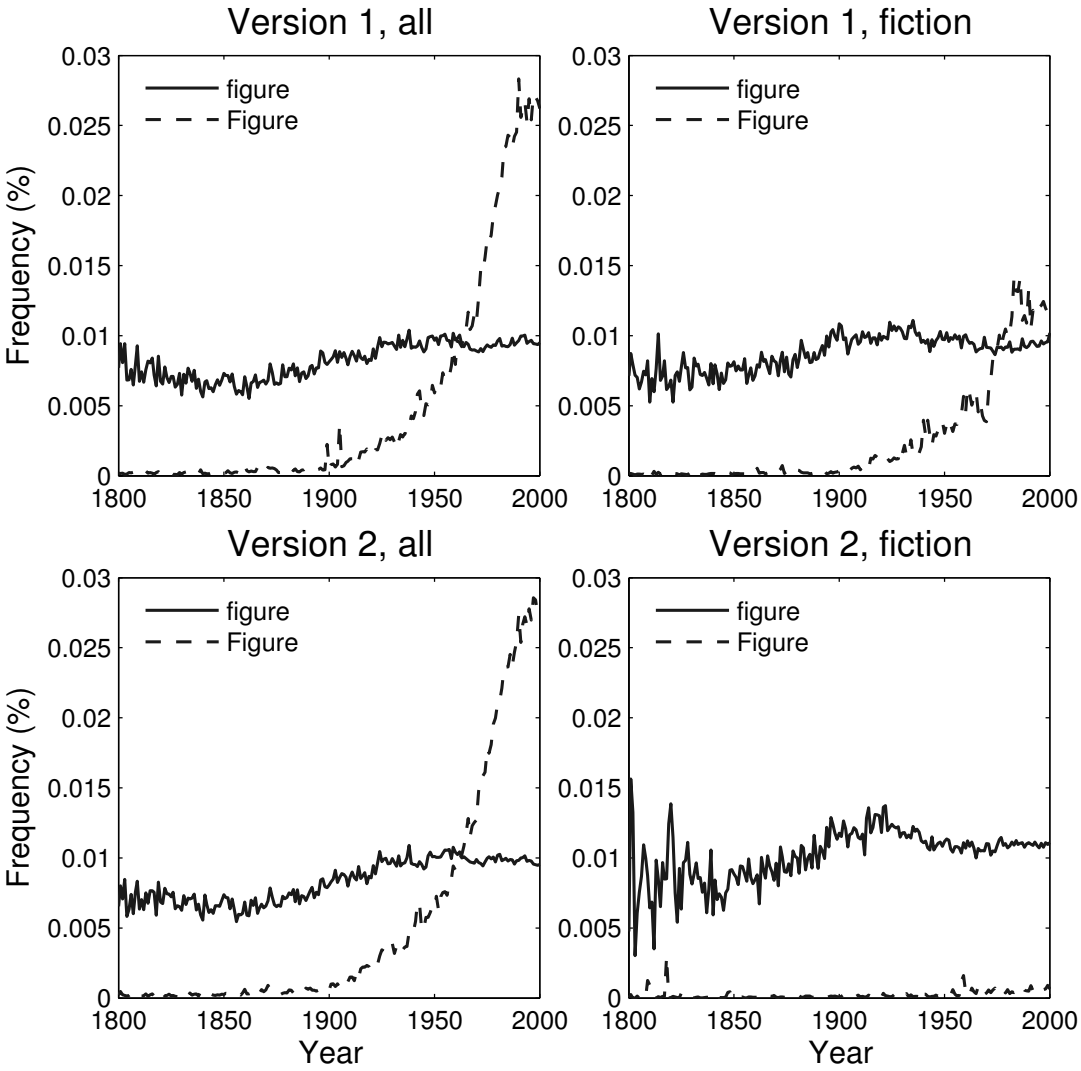

Figure 2.2: Relative frequencies of "Figure" vs "figure" in both versions of the Google Books corpus for both English (all) and English Fiction. In the English data sets, the capitalized term rapidly surpasses the uncapitalized term in the 1960s. For the first English Fiction data set, this effect is delayed until the 1970s. This effect is only avoided in the second version of the English Fiction data set. These trends strongly suggest an increase in the sampling of professional texts to both English data sets and the first English Fiction data set over time.

2.2 Methods

2.2.1 Statistical divergence between years

We examine the dynamics of a data set by calculating the statistical divergence between the distributions of 1-grams in two given years. A commonly used measure of statistical divergence is Kullback-Leibler (KL) divergence [11], based on which we use a bounded, symmetric measure. Given a language with N unique words and 1-gram distributions P in the first year and Q in second, the KL divergence between P and Q can be expressed as

$$D_{KL}(P||Q) = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}, \tag{2.1}$$

where p_i is the probability of observing the i^{th} 1-gram in a random text from the distribution for first year, and q_i is the probability of observing the same word in an analogous text for the second year. If the base of the logarithm is two, then divergence has a unit of bits; moreover, it may be interpreted as the average number of bits wasted if a text from the first year is encoded efficiently, but according to the distribution from the latter, incorrect year. To demonstrate this, we may rewrite the previous equation as

$$D_{KL}(P||Q) = -\sum_{i=1}^{N} q_i \log p_i - H(P), \tag{2.2}$$

where $H(P) = -\sum_i p_i \log p_i$ is the Shannon entropy [12], also the average number of bits required per word in an efficient encoding for the original distribution; and the remaining term is the average number of bits required per word in an efficient, but

mistaken, encoding of a given text. However, if a single (say, the j^{th}) 1-gram in the language exists in the first year, but not in the second, then $q_j = 0$, and the divergence diverges. Since this scenario is not extraordinary for the data sets in question, we instead use Jensen-Shannon divergence (JSD) [13] given by

$$D_{JS}(P||Q) = \frac{1}{2}\Big(D_{KL}(P||M) + D_{KL}(Q||M)\Big), \tag{2.3}$$

where $M = \frac{1}{2}(P + Q)$ is a mixed distribution of the two years. This measure of divergence is bounded between 0 when the distributions are the same and 1 bit in the extreme case when there is no overlap between the 1-grams in the two distributions. In fact, if we begin with a uniform distribution of N species and replace k of those species with k entirely new ones, the JSD between the original and new distribution is k/N, the proportion of species replaced. The JSD is also symmetric, which is an added convenience. The JSD may be expressed as

$$D_{JS}(P||Q) = H(M) - \frac{1}{2}\Big(H(P) + H(Q)\Big), \tag{2.4}$$

from which it is apparent that a similar waste analogy holds as with KL divergence, with the mixed distribution taking the place of the approximation regardless of the year a text was written.

2.2.2 Key contributions of individivual words

The form for Jensen-Shannon divergence given in Eq. 2.4 can be broken down into contributions from individual words, where the contribution from the i^{th} word to the

divergence between two years is given by

$$D_{JS,i}(P||Q) = -m_i \log m_i + \frac{1}{2}\Big(p_i \log p_i + q_i \log q_i\Big). \tag{2.5}$$

Some rearragement gives

$$D_{JS,i}(P||Q) = m_i \cdot \frac{1}{2}\Big(r_i \log r_i + (2 - r_i) \log(2 - r_i)\Big), \tag{2.6}$$

where $r_i = p_i/m_i$, so that contribution from an individual word is proportional to the average frequency of the word, and the proportion depends on the ratio between the smaller frequency (without loss of generality) and the average. Namely, we may reframe the equation above as

$$D_{JS,i}(P||Q) = m_i C(r_i). \tag{2.7}$$

Words with larger average frequency yield greater contributions as do those with smaller ratios, r, between the smaller and average frequency. So while a common 1-gram—such as "the," "if," or a period—changing subtly can have a large effect on the divergence, so can an uncommon (or entirely new) word given a sufficient shift from one year to the next. The size of the contribution relative to the average frequency is displayed in Fig. 2.3 for ratios ranging from 0 to 1. $C(r_i)$ is symmetric about $r_i = 1$ (i.e., no change), so no novel behavior is lost by omitting the case where $r_i > 1$ (i.e., when p_i is the larger frequency). The maximum possible contribution (in bits) is precisely the average frequency of the word in question, which occurs if and only

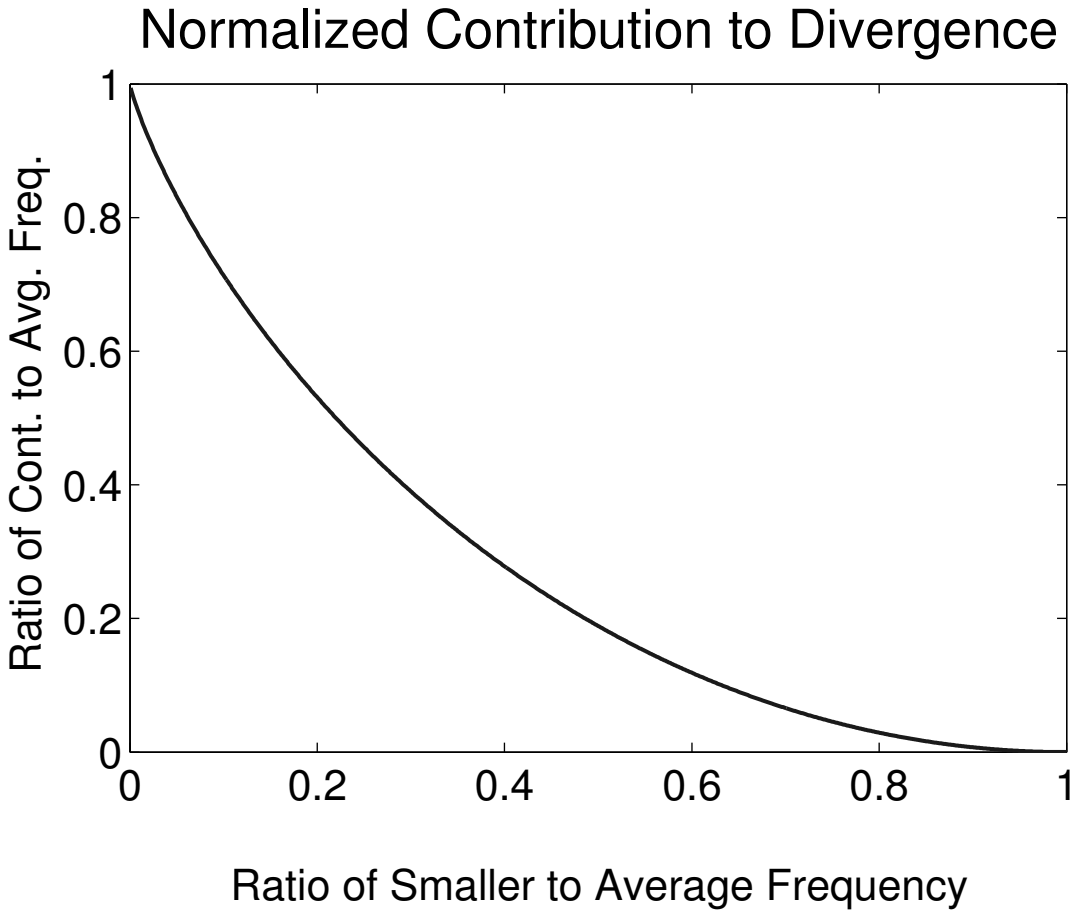

Figure 2.3: For the ratio r between the smaller relative frequency of an element and the average, $C(r)$ is the proportion of the average contributed to the Jensen-Shannon divergence (see Eqs. 2.6 and 2.7). In particular, if $r = 1$ (no change), then the contribution is zero; if $r = 0$, the contribution is half its frequency in the distribution in which it occurs with nonzero frequency.

if the smaller frequency is 0. No contribution is made if and only if the frequency remains unchanged.

We course-grain the data at the level of decades—e.g., between 1800-to-1809 and 1990-to-1999—by averaging the relative frequency of each unique word in a given decade over all years in that decade. (Each year is weighted equally.) This allows

convenient calculation and sorting of contributions to divergence of individual 1-grams between any two time periods.

2.3 Results and Discussion

Fig. 2.4 shows the JSD between the 1-gram distributions for every pair of years between 1800 and 2000 contributed by 1-grams present above a threshold normalized frequency of 10^{-5} for both versions of the English and English Fiction data sets (i.e., words that appear with frequency at least 1 in 10^5). A major qualitative aspect apparent from the heatmaps is a gradual increase in divergence with differences in time. However, there are several exceptions to this rule of thumb. First among these are two cross-hair patterns, where the image is "pinched" toward the diagonal, in the vicinities of the two world wars. Also visible is an asymmetry that suggests a particularly high divergence between the first half century and the last quarter century observed. We examine these effects more closely in Figs. 2.5 and 2.6 by taking specific slices of the heatmaps. In particular, we consider the divergences of each year from 1880 by examining the appropriate row—or column, equivalently—and the divergences between consecutive years by way of the off-diagonal. (Also included to verify qualitative consistency are the analogous contribution curves using the more restrictive threshold of 10^{-4}.) While the initial divergence between any two consecutive years is noticeable, the divergence increases (for the most part) steadily with the time difference. The cross-hairs from the heatmap resolve into war-time bumps in divergence, which quickly settle in peacetime. The larger boost to the divergence in recent decades, however, is more persistent suggesting a more fundamental change in

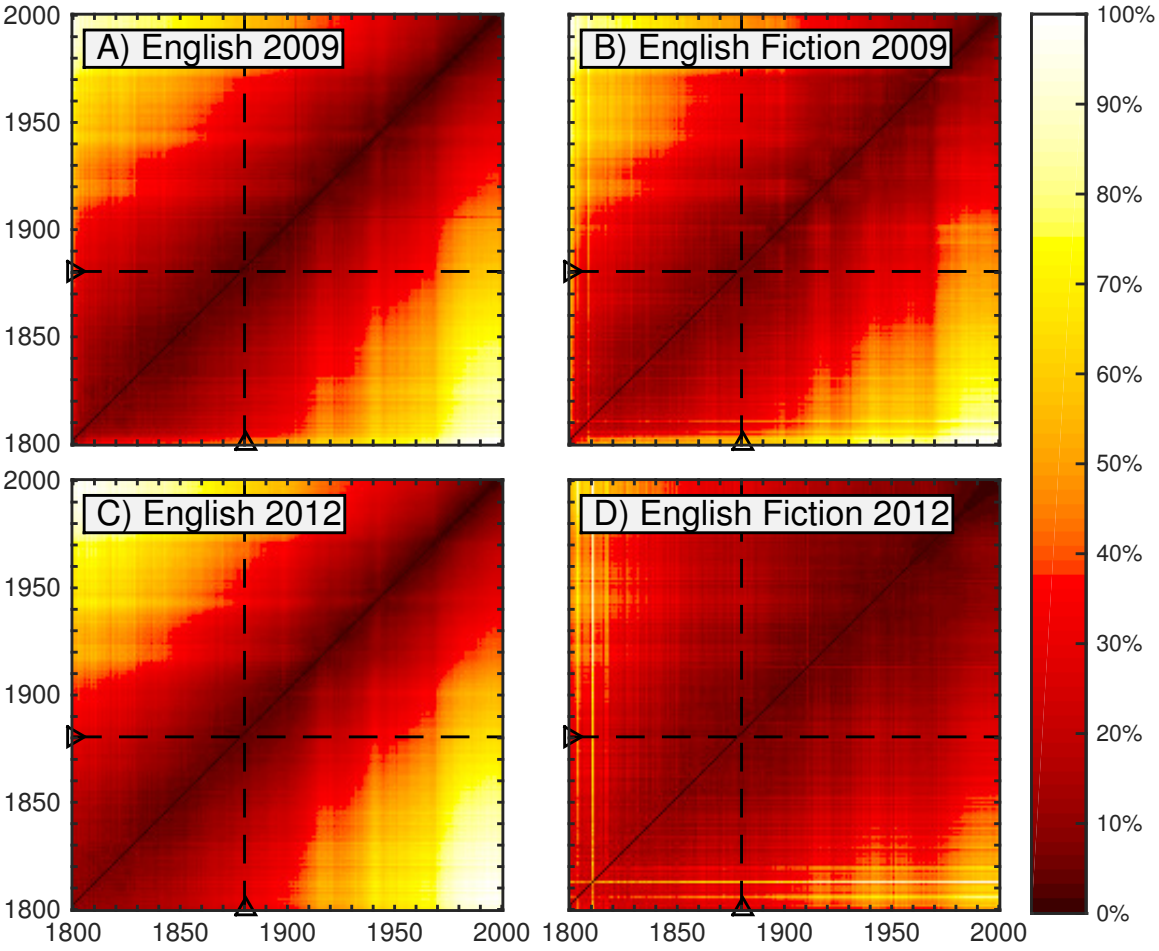

Figure 2.4: Heatmaps showing the JSD between every pair of years between 1800 and 2000, contributed by words appearing above a frequency threshold of 10^{-5}. The dashed lines highlight the divergences to and from the year 1880, which are featured in Fig. 2.5. The off-diagonal elements represent divergences between consecutive years, as in Fig. 2.6. The color represents the percentage of the maximum divergence observed in the given time range for each data set. The divergence between a year and itself is zero. For any given year, the divergence increases with the distance (number of years) from the diagonal—sharply at first, then gradually. Interesting features of the maps are the presence of two cross-hairs in the first half of the 20th century, which strongly suggests a wartime shift in the language, as well as an asymmetry that suggests a particularly high divergence between the first half century and the last quarter century observed.

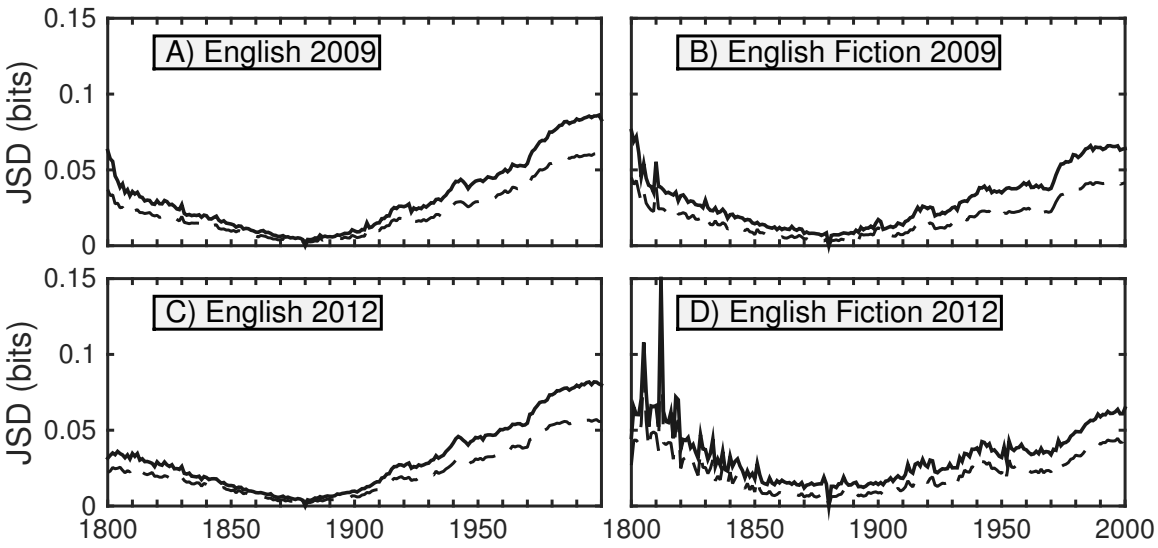

Figure 2.5: JSD between 1880 and each displayed year for given data set, corresponding to dashed lines from Fig. 2.4. Contributions are counted for all words appearing above a 10^{-5} threshold in a given year; for the dashed curves, the threshold is 10^{-4}. Typical behavior in each case consists of a relatively large jump between one year and the next with a more gradual rise afterward (in both directions). Exceptions include wartime, particularly the two World Wars, during which the divergence is greater than usual; however, after the conclusion of these periods, the cumulative divergence settles back to the previous trend. Initial spikiness in (D) is likely due to low volume.

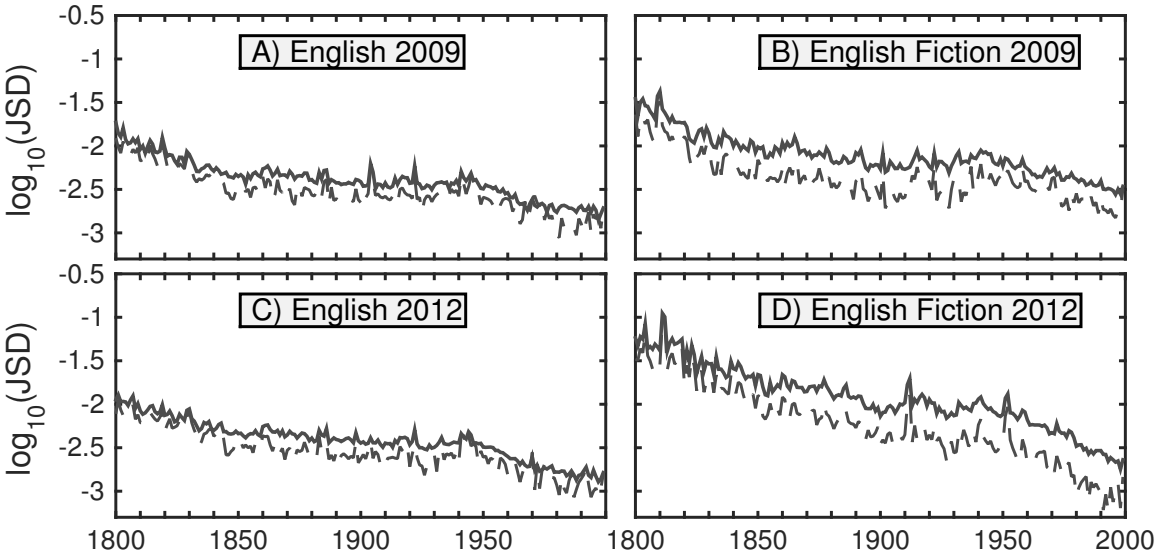

Figure 2.6: Consecutive year (between each year and the following year) base-10 logarithms of JSD, corresponding to off-diagonals in Fig. 2.4. For the solid curves, contributions are counted for all words appearing above a 10^{-5} threshold in a given year; for the dashed curves, the threshold is 10^{-4}. Divergences between consecutive years typically decline through the mid-19th century, remain relatively steady until the mid-20th century, then continue to decline gradually over time.

the data set, which we will examine in more depth later in this section. Divergences between consecutive years typically decline through the mid-19th century. Divergences then remain relatively steady until the mid-20th century, then continue to decline gradually over time, which may be consistent with previous findings of decreased rates of word introduction and increased rates of word obsolescence in many Google Books data sets over time [6] and a slowing down of linguistic evolution over time as the vocabulary of a language expands [8]. The initial spikes in divergence in the second version of the fiction data set are likely due to the lower initial volume observed in Fig. 2.1.

Selected examples of the top 60 contributions to inter-decade divergence are given in Figs. 2.7 to 2.12. (For the rest, see the Supporting Online Materials [14].) The largest contributions to all divergences generally appear to be from increased relative frequencies of use of words between decades. (We will examine this apparent tendency more closely in a future paper.) For the unfiltered data, these are in turn heavily influenced by increased mention of years. Divergences in English Fiction are not strongly affected by years (also see Fig. 2.15). The 1940s literature, unsurprisingly, features more references to Hitler and war than the 1930s, along with other World War II-related military and political terms. This is seen regardless of the specific data set used and is fairly encouraging. Curiously, regardless of the specific data set, a noticeable contribution is given by an increase in relative use of the words "Lanny" and "Budd," in reference to one character (Lanny Budd) frequently written about by Upton Sinclair during that decade. In fact, in the fiction data sets, this character dominates the charts.

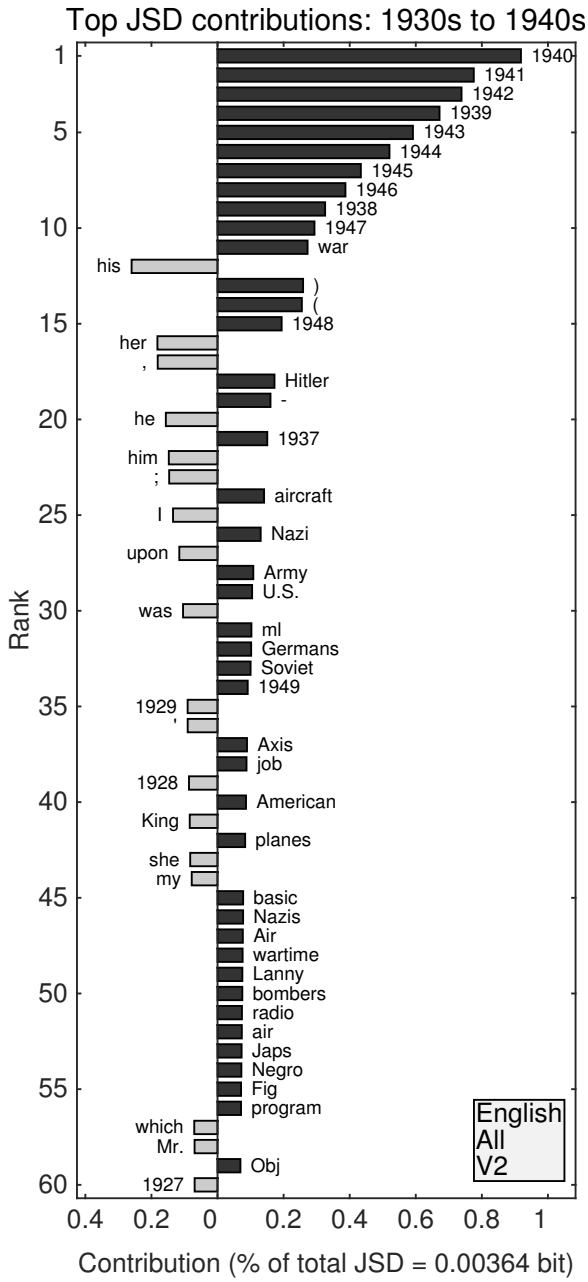

Figure 2.7: (English, all; Version 2.) Top 60 individual contributions of 1-grams to the JSD between the 1930s and the 1940s. Each contribution is given as a percentage of the total JSD (see horizontal axis label) between the two given decades. All contributions are positive; bars to the left of center represent words that were more common in the earlier decade, whereas bars to the right represent words that became more common in the later decade.

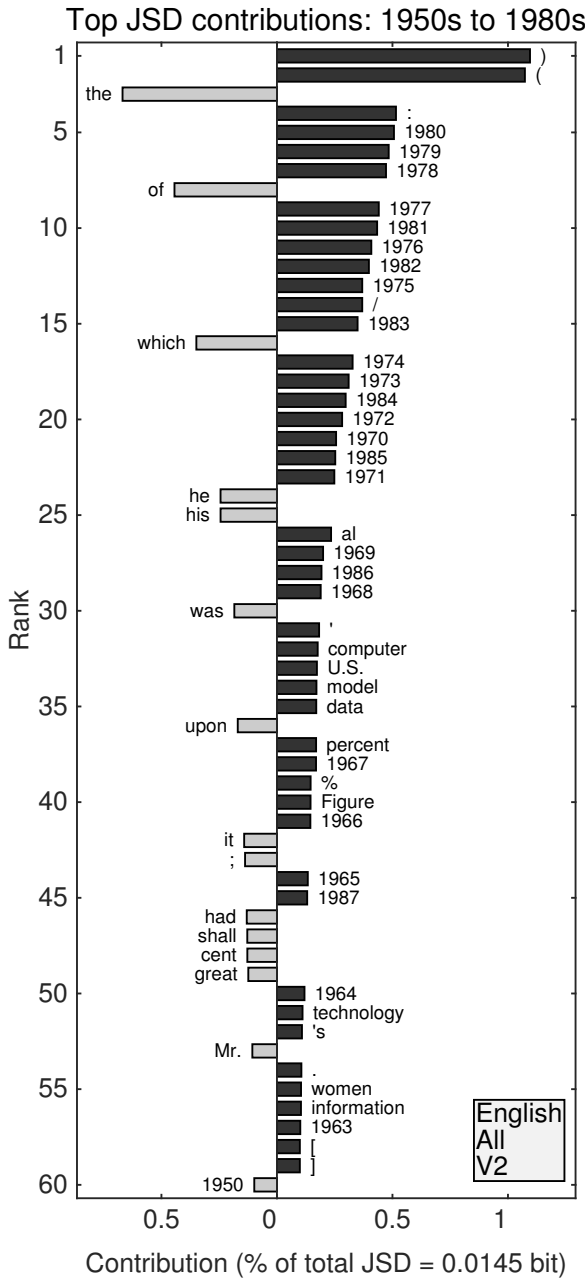

Figure 2.8: (English, all; Version 2.) Top 60 individual contributions of 1-grams to the JSD between the 1950s and the 1980s. Each contribution is given as a percentage of the total JSD (see horizontal axis label) between the two given decades. All contributions are positive; bars to the left of center represent words that were more common in the earlier decade, whereas bars to the right represent words that became more common in the later decade.

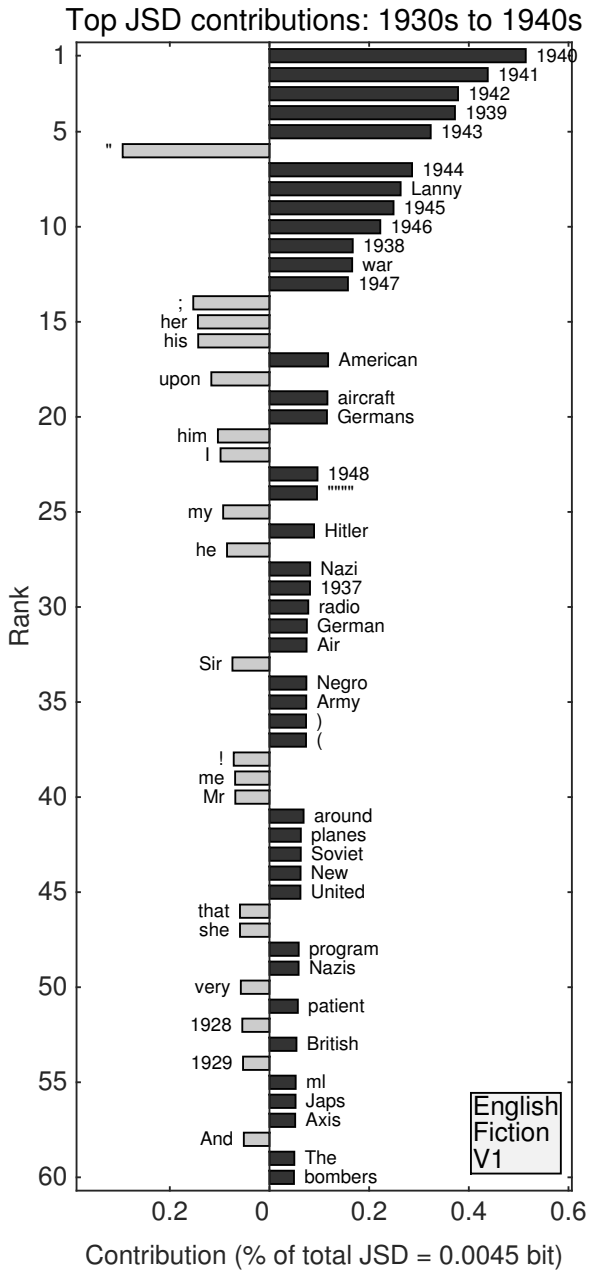

Figure 2.9: (English Fiction, Version 1.) Top 60 individual contributions of 1-grams to the JSD between the 1930s and the 1940s. Each contribution is given as a percentage of the total JSD (see horizontal axis label) between the two given decades. All contributions are positive; bars to the left of center represent words that were more common in the earlier decade, whereas bars to the right represent words that became more common in the later decade.

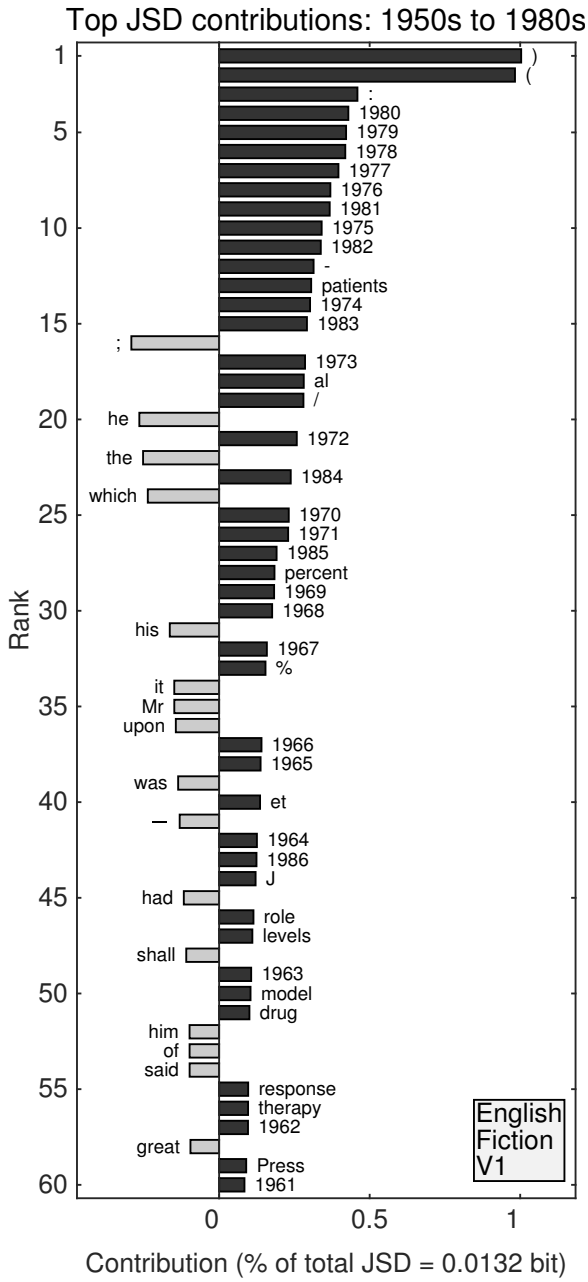

Figure 2.10: (English Fiction, Version 1.) Top 60 individual contributions of 1-grams to the JSD between the 1950s and the 1980s. Each contribution is given as a percentage of the total JSD (see horizontal axis label) between the two given decades. All contributions are positive; bars to the left of center represent words that were more common in the earlier decade, whereas bars to the right represent words that became more common in the later decade.

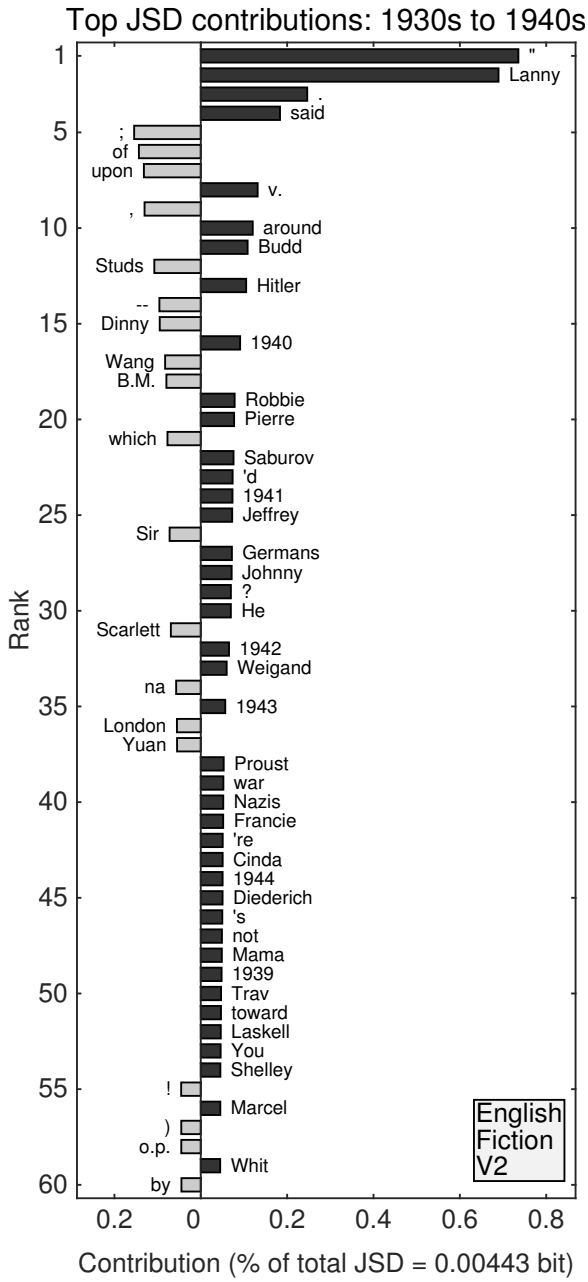

Figure 2.11: (English Fiction, Version 2.) Top 60 individual contributions of 1-grams to the JSD between the 1930s and the 1940s. Each contribution is given as a percentage of the total JSD (see horizontal axis label) between the two given decades. All contributions are positive; bars to the left of center represent words that were more common in the earlier decade, whereas bars to the right represent words that became more common in the later decade.

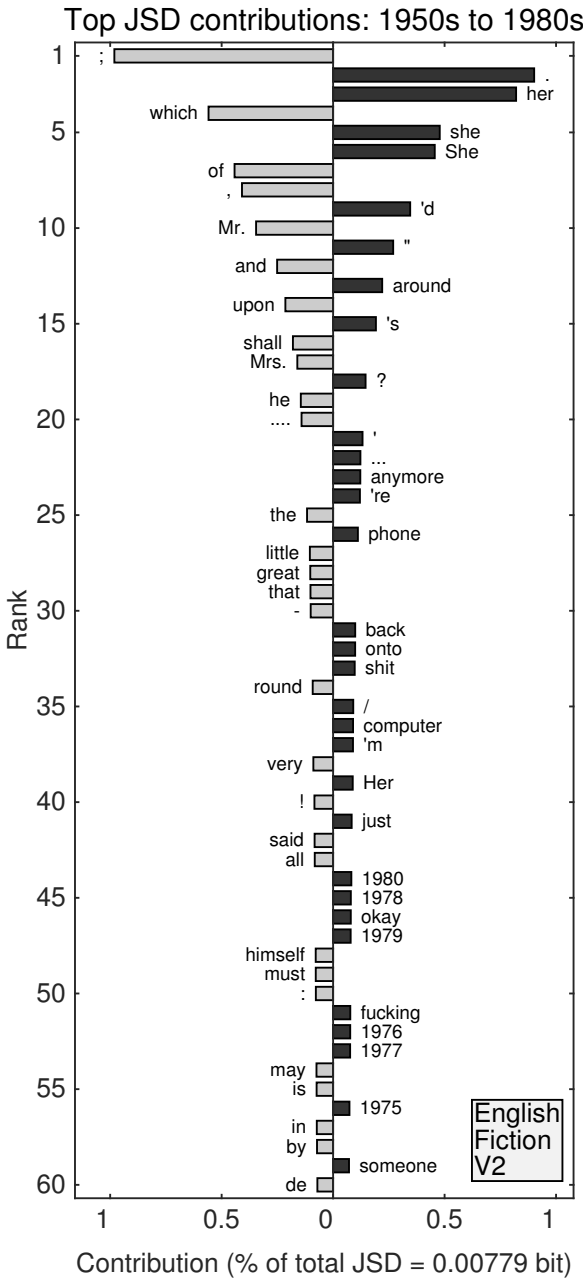

Figure 2.12: (English Fiction, Version 2.) Top 60 individual contributions of 1-grams to the JSD between the 1950s and the 1980s. Each contribution is given as a percentage of the total JSD (see horizontal axis label) between the two given decades (see title). All contributions are positive; bars to the left of center represent words that were more common in the earlier decade, whereas bars to the right represent words that became more common in the later decade.

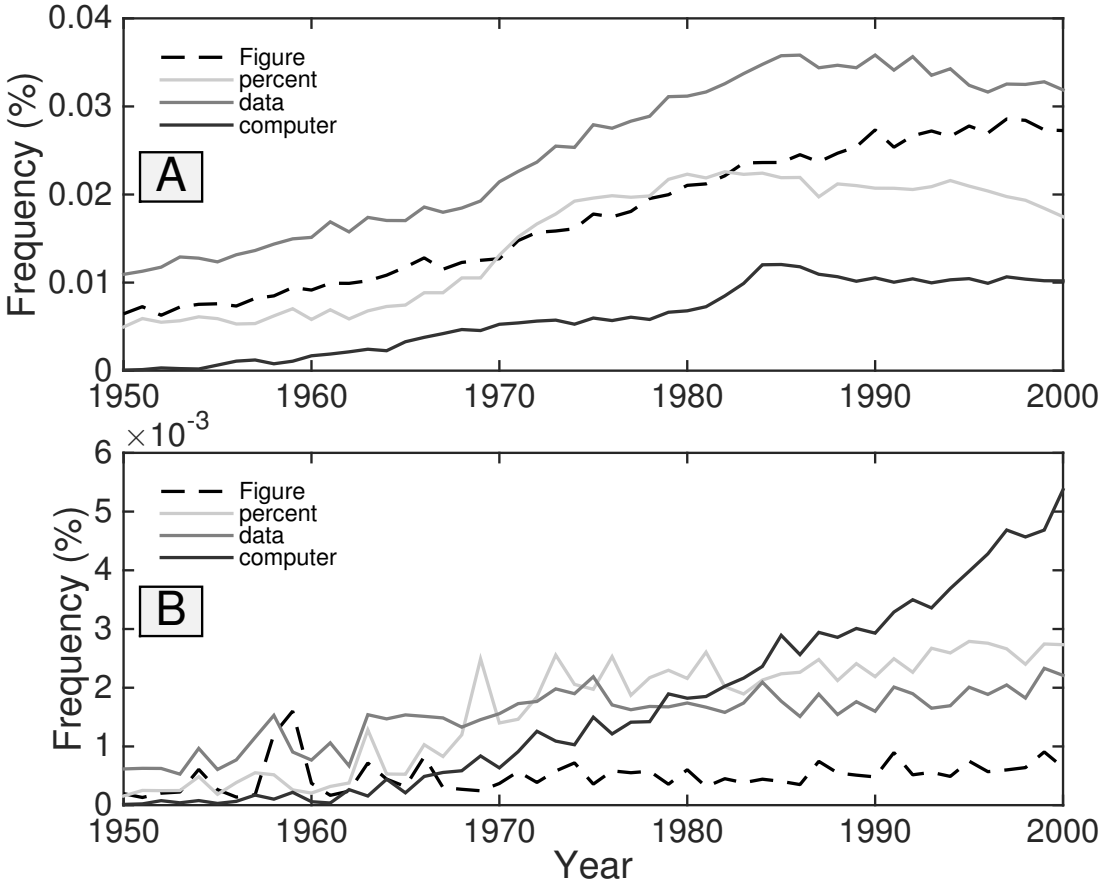

Figure 2.13: Time series of technical terms from Version 2: (a) English all, (b) English fiction. In the unfiltered data set, these technical terms appear frequently and increase in usage though the 1980s. In fiction, technical terms show up far less frequently and remain relatively stable in usage with the notable exception of "computer," which has been gradually gaining popularity since the 1960s.

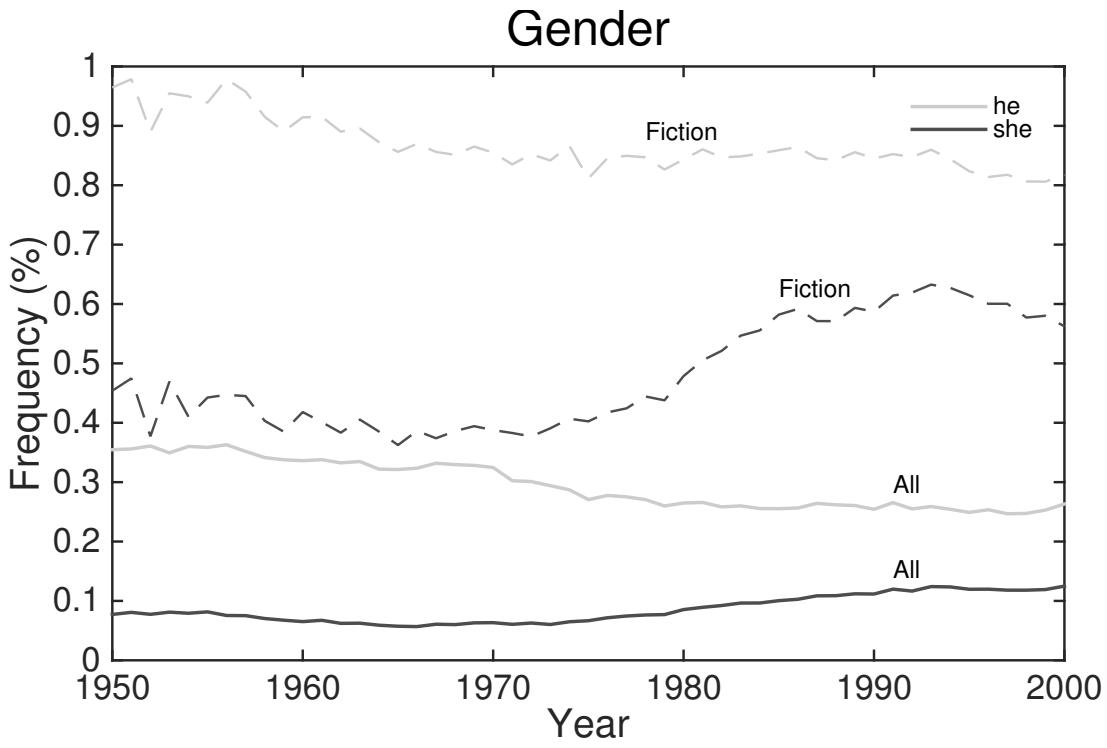

Figure 2.14: Time series for "he" and "she" for Version 2. The unfiltered frequencies are given by the solid curve. Frequencies in fiction are given by the dashed curve. These personal pronouns are more common in fiction. The pronoun "she" gains popularity through the 1990s in both data sets; however, this effect is more pronounced in fiction.

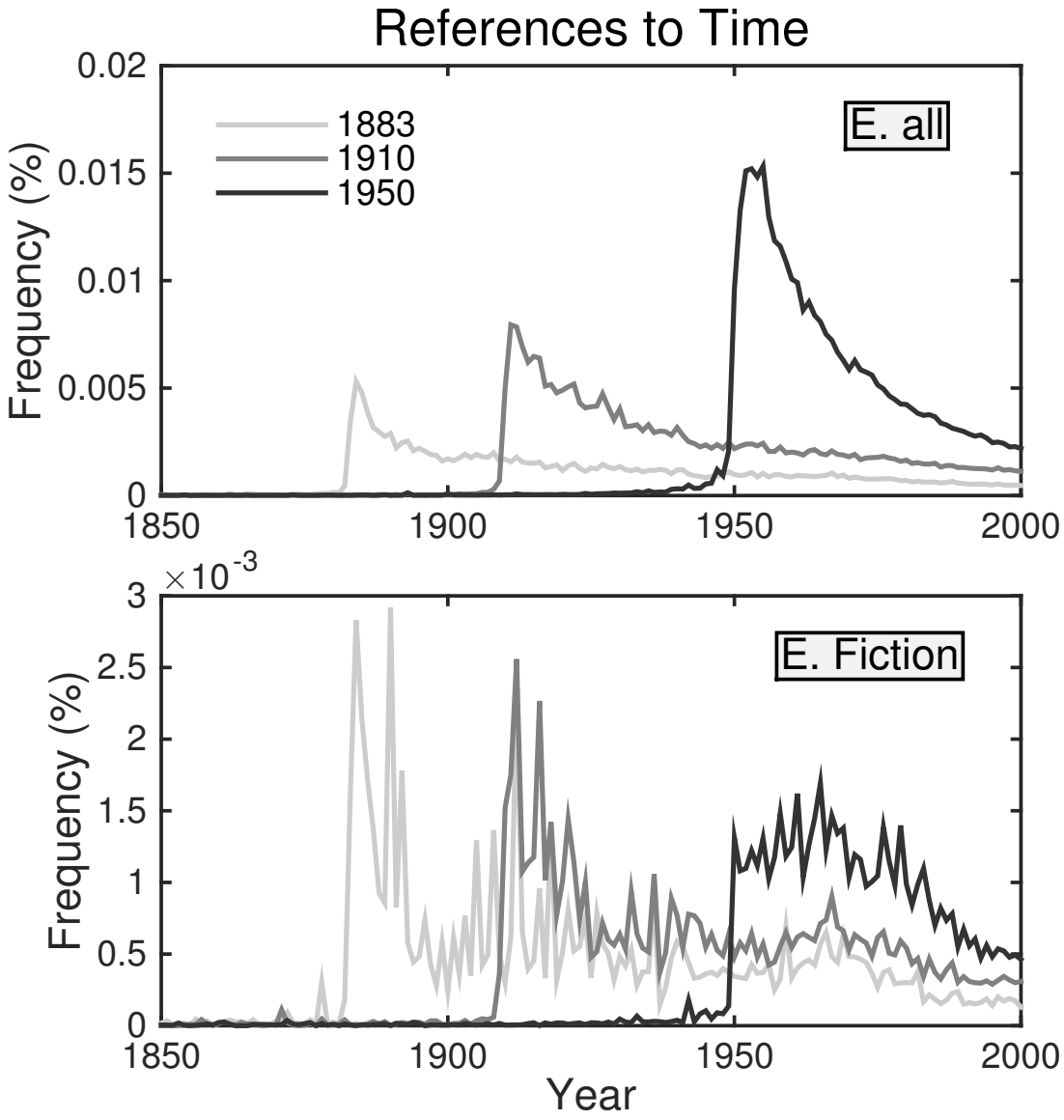

Figure 2.15: Frequencies of references to years. Top deliberately resembles a figure from [1] using unfiltered data from English Version 2. (The cited paper uses Version 1.) Note the characteristic rapid rises and gradual declines, as well as the increasing peaks in yearly references. However, while the characteristic shape is still present in fiction (Version 2, bottom)—at much reduced levels—the peaks do not rise. The rising effect is likely due to citations from professional texts.

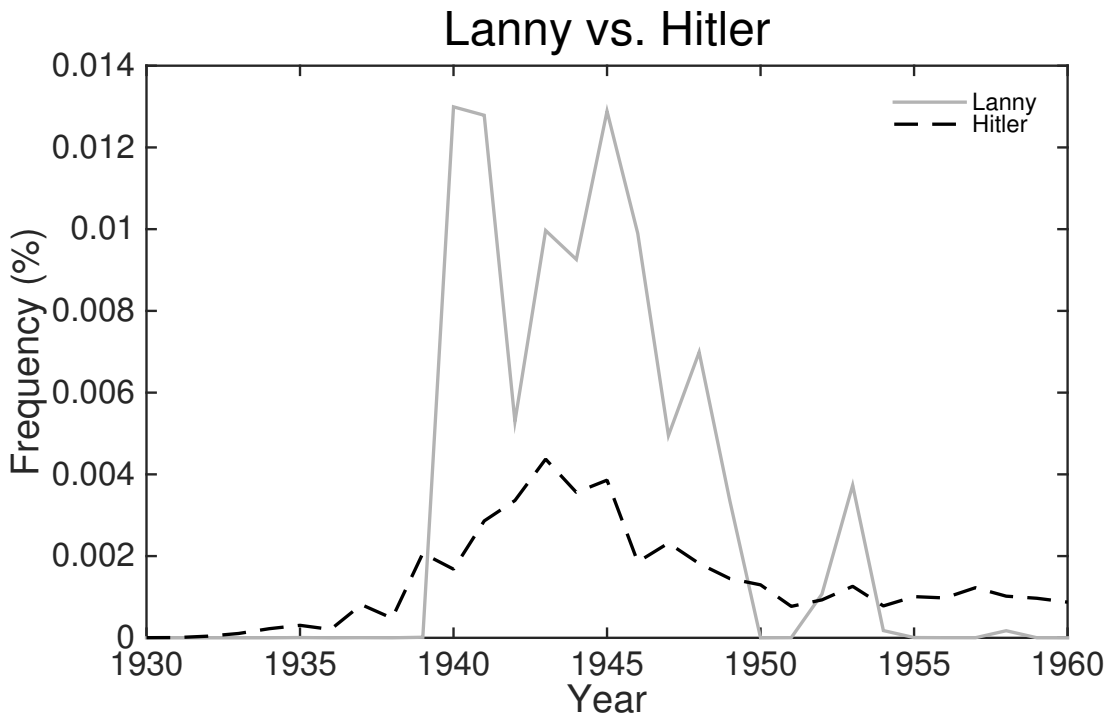

Figure 2.16: Upton Sinclair wrote 11 Lanny Budd novels set during World War II. The first of these was published in 1940, and the last was published in 1953. The net effect of Sinclair's efforts is that his character appears a lot more frequently in the English Fiction (Version 2) data set than Hitler during most of the war. This demonstrates the potential impact of a single prolific author on the corpus.

The second version of the unfiltered English data set in the 1930s and 1940s (see Fig. 2.7) has dynamics dominated by references to years. (The first version is similar. For analogous figures see the Supporting Online Materials [14].) In fact, eight of the top ten contributions to the divergence between those decades are due to increased relative frequencies of use of each of years between 1940 and 1949. The other two top ten words are the last two years of the previous decade, which also increased in relative frequency of use. ("1948" and "1949" appear at ranks 15 and 34, respectively.) The last three years in the 1920s also appear by way of decreased relative frequency of use in the top 60 contributions. The 11th highest contribution is from "war," which increased in relative frequency. "Hitler" and "Nazi" (increased relative frequencies) are ranked 18th and 26th, respectively. Parentheses (13th and 14th) show increased relative frequencies of use. Personal pronouns show decreased relative frequencies of use. The word "King" (41st) also shows a decreased relative frequency, possibly due to the British line of succession.

The top two contributions between the 1950s and the 1980s (see Fig. 2.8) in the English data set are both parentheses, which show dramatically increased relative frequencies of use. Combined with increased relative frequencies for the colon (4th), forward slash (14th), "computer" (32nd), and square brackets (58th and 59th), this suggests that the primary changes between the 1950s and the 1980s are due specifically to computational sources. Other technical words showing noticeable increases include "model" (34th), "data" (35th), "percent" and the percentage sign (37th and 39th), "Figure" (40th), "technology" (51st), and "information" (56th). Similarly to the divergence between the 1930s and 1940s, 19 out of the top 30 places are accounted for by increased relative frequencies of use in years between 1968 and 1980. The

words “the” (3rd), “of” (8th), and “which” (16th) all decrease noticeably in relative frequency and are the highest ranked alphabetical 1-grams. Unlike the divergence between the 1930s and 1940s, only masculine pronouns show decreases in the top 60. In fact, “women” (55th) increases.

The first version of English Fiction shows similar dynamics to the second version of the unfiltered data set between the 1930s and the 1940s (see Fig. 2.9) with yearly mentions dominating the ranks. Some exceptions include “Lanny” rising in rank from 49th to 8th, parentheses falling from 13th and 14th to 36th and 37th, and “ml” (increased relative frequency of use in the 1940s) falling from 31st to 55th, and “radio” (with increased relative frequency) rising from 51st to 30th. Moreover, while “King” is no longer in the top 60 contributions, “patient” (ranked 51st) is with an increased relative frequency of use in the 1940s.

This similarity between the original English Fiction data set and the unfiltered data set also appears in the divergence between the 1950s and the 1980s (see Fig. 2.10) with parentheses and years dominating. Moreover, “patients” ranks 13th (with increased relative frequency of use) despite not appearing in the top 60 for the unfiltered data set. These observations, combined with increases in “levels” (47th), “drug” (51st), “response” (55th), and “therapy” (56th) demonstrate the original fiction data set is strongly influenced by medical journals. Therefore, this data set cannot be considered as primarily fiction despite the label.

Fortunately, the same is not true for the second version of the English Fiction data set. This is quickly apparent upon inspection of the two greatest contributions to the divergence between the 1930s and the 1940s (see Fig. 2.11). The first of these is due to a dramatic increase in the relative frequencies of use of quotation marks, which in

turn implies increased dialogue. The second is the name "Lanny" in reference to the recurring character Lanny Budd from 11 Upton Sinclair novels published between 1940 and 1953. "Budd" ranks 11th in the chart ahead of "Hitler" (13th). The frequency series for "Lanny" and "Hitler" provided in Fig. 2.16 demonstrate that Lanny, in fact, received more mention than Hitler during this time period. In fact, the chart is littered with the names of fictional characters. Studs Lonigan, the 1930s protagonist of a James T. Farrel trilogy, secures the 12th spot. (Naturally, he is mentioned fewer times during the 1940s.) Dinny Cherrel from the 1930s *The Forsyte Saga* by John Galsworthy secures rank 15. Wang Yuan from the 1930s *The House of Earth* trilogy by Pearl S. Buck ranks 17th and 37th. Detective Bill Weigand, a recurring character created by Richard Lockridge in the 1940s, secures rank 33. The eponymous, original Asimov robot from the 1940 short story, "Robbie," ranks 19th. "Mama" (ranked 48th) is none other than the subject of *Mama's Bank Account*, published in 1943 by Kathryn Forbes. "Saburov" (ranked 22nd) from *Days and Nights* by Konstantin Simonov and "Diederich" (ranked 45th) from *Der Untertan* by Heinrich Mann are subjects of works translated into English in the 1940s. So while Marcel Proust (56th and 33rd) who died in 1922 may be present in the 1940s due to letters translated by Mina Curtiss in 1949 or other references not technically fiction—similarly, "B.M." (18th) may refer to the author B. M. Bower—the vast majority of prominent words in this chart may be traced not only to authors of fiction, but to the content of their work. Moreover, the greatest contributions to divergence appear to correspond to the most prolific authors, particularly Upton Sinclair.

While there are no names of characters in the top divergences between the 1950s and the 1980s, the updated fiction data set (Fig. 2.12) displays far more variety

than the original version, including decreases in relative frequencies of masculine pronouns—e.g., "he" (rank 19) and "himself" (rank 48)—and corresponding increases for feminine pronouns—e.g., "her" (3rd), "she" (5th), and "She" (6th)—(also see Fig. 2.14), an increase in relative frequencies of contractions (see ranks 9, 15, and 21), a decrease in "shall" (16th) and "must" (49th), and a variety of increased profanity (particularly ranks 33 and 51). "Mr." (10th) and "Mrs." (17th) both see decreased relative frequencies of use. Various shifts in punctuation are present, particularly fewer semicolons (1st) and more periods (2nd). Quotation (11th) and question (18th) marks both see increased relative frequencies of use in the 1980s, and the four-period ellipsis (20th) loses ground to the three-period version (22nd). The word "computer" is also more common in the 1980s. In fact, as shown in Fig. 2.13, "computer" gains popularity in the fiction data set despite other technical words remaining relatively steady in usage. This picture of the second fiction data set should be encouraging for anyone attempting to analyze colloquial English, despite the prolificity bias apparent from the effects Sinclair and other authors had on the divergence between the 1930s and 1940s.

In the Supporting Online Material [14], we include the top 60 contributions to divergences between each pair of the 20 decades in each of the four data sets analyzed in this paper. In total, 760 figures are included (190 per data set) for a grand total of 45,600 contributions. We highlight some of these here.

- For divergences to and from the first decade of the 1800s, many of the contributions are due to a reduction of optical character recognition confusion between the letters 'f' and 's'. For example, in the second unfiltered data set between the 1800s and 1810s, the top two contributions are due to reductions in "fame" and

"os," respectively. The word "same" (ranked 11th) is the first increasing contribution. Decreased relative frequencies of "os," "ï£ijsirst," "thofe," "fo," "fay," "cafe," "fays," "fome," and "faid" (ranks 3 through 10, respectively) and "lise" (12th) all suggest digital misreadings of both 'f' and the long 's'. (The 13th contribution is "Napoleon," who is mentioned with greater relative frequently in the 1810s.)

- Contributions between the 1830s and the 1860s in the second unfiltered data set highlight the American Civil War and its aftermath. "State" (11th), "General" (19th), "States" (20th), "Union" (37th), "Confederate" (48th), "Government" (52nd), "Federal" (56th), and "Constitution" (59th) all show increased relative frequency of use. Religious terms tend to decline during this period—e.g., "church" (14th), "God" (24th), and "religion" (58th).

- Between the 1940s and 1960s, the second unfiltered dataset shows increases for "nuclear" (43rd), "Vietnam" (47th), and "Communist" (50th). The relative frequency of "war" (25th) decreases substantially. Meanwhile in fiction, "Lanny" (5th) declines, while "television" (38th) and the Hardy Boys ("Hardy" ranks 51st) appear with greater relative frequencies.

- Between the 1960s and 1970s, the second fiction data set is strongly affected by "Garp" (*The World According to Garp* by John Irving, 1978) at rank 19, increased relative frequencies of profanity (ranks 27, 33, and 38), and increased mentions of "Nixon" (41st) and "Spock" (47th, likely due to "Star Trek" novels).

- Between the 1980s and 1990s, the second fiction set shows increased relative frequencies of use of the words "gay" (15th), "lesbian" (19th), "AIDS" (24th),

and "gender" (27th). Female pronouns (2nd, 8th, and 9th) show increased relative frequencies of use in continuance of Fig. 2.12.

2.4 Concluding remarks

The unfiltered English data sets are similar between versions and appear to be dominated by an increase in references to recent years, as does the first fiction data set. In fact, there appears to be a general asymmetry in the contributions to divergence with more being due to relative frequencies increasing with time than to relative frequencies decreasing with time. We will examine this apparent asymmetry more closely in a future paper.

The unfiltered data sets feature more general terms such as "percent," "data," "Figure," and "model." (Also see Fig. 2.13.) The original fiction data set also features these, but also places "patients," "drug," "response," and "therapy" among the top 60 contributions. In fact, the primary difference between the unfiltered and original fiction data sets in the 1980s (compared to the 1950s) appears to consist of the nature of journals sampled. The unfiltered components predicted and observed for this particular data set seem to be dominated by medical journals.

As well as having more mentions of time and technical terms (and parentheses) in the 1980s than in the 1950s, both unfiltered versions and the first fiction data set include both "et" and "al" with greater relative frequency in the 1980s. Perhaps more importantly, years do not have a large effect on the dynamics in the second English Fiction data set. In fact, we see in Fig. 2.15 that while peaks for years rise in the unfiltered data, they do not in fiction. The absence of rising peaks in fiction strongly

suggests the rise in peak relative frequencies of years in the larger data set is due to a citation bias in the unfiltered data set from high sampling of scientific journals. This bias casts doubt on conclusions that we as a culture forget things more quickly than we once did based on the observation that half-lives for mentions of a given year decline over time [1].

The exponentional rise in scientific literature is not a new phenomenon, and as Derek John de Solla Price stated 51 years ago [15] (p. 81) when discussing the half-lives for citations of scientific literature, "In fields embarrassed by an inundation of literature there will be a tendency to bury as much of the past as possible and to cite older papers less often than is their statistical due." In short, it may very well be that an explanation for declining half-lives in the mentions of years need not invoke an evolution of cultural memory.

A plausible strategy when dealing with a new, large data set is to begin by testing familiar ground to determine if analysis of the data yields familiar results. Then, when it does, one can procede to search for novel results. If we expect to see masculine pronouns in decline, and then we really do see this along with other familiar patterns in this data set, then we would like to lend credence to observations regarding the alteration of growth and death rates of words, the public perception of time, and so forth. However, it is clear from this analysis that the contents of the Google Books data sets do not represent an unbiased sampling of publications, especially in recent decades in which change appears dominated by professional publications rather than popular works to the point that even the first data set specifically labeled as fiction appears dominated by medical literature. Therefore, it is necessary when examining these data sets to quantify the popular and professional components in order to form

a more socially accurate picture of the corpus. For instance, one should ask how much of any observed gender shift is due to shifts in popular works and how much is due to changes in professional norms, as well as which one preceeds the other. (In the case of gender pronouns, we actually do observe periods of feminization in the second version of the fiction data set, which may support related findings.)

Even where popular works are involved, the frequencies of words are not a direct measure of popularity of those words. Not only can there be a delay between popularities of words and their publication, frequencies may also capture the prolificity of the authors using those words rather than their popularity to the general public. In the case of Lanny Budd, a character was vaulted (above Hitler) to the upper echelon of words affecting divergence by virtue of Upton Sinclair writing 11 novels featuring this one character between 1940 and 1953. This effect further establishes the data as a proxy for social information after the fact.

Therefore, while the size, availability, and the power of Google Books to highlight numerous trends is beyond doubt, a cautious approach is necessary for any attempt to produce novel results. Our analysis provides a framework for improvements to previous and future works which, if performed on English data, ought to validate results with the 2012 version of the English Fiction data set or otherwise account for the biases of the corpus.

References

[1] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

REFERENCES

[2] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the google books ngram corpus," in *Proceedings of the ACL 2012 System Demonstrations*, pp. 169–174, Association for Computational Linguistics, 2012.

[3] J. M. Twenge, W. K. Campbell, and B. Gentile, "Increases in individualistic words and phrases in american books, 1960–2008," *PloS one*, vol. 7, no. 7, 2012.

[4] J. M. Twenge, W. K. Campbell, and B. Gentile, "Male and female pronoun use in us books reflects women's status, 1900–2008," *Sex roles*, vol. 67, no. 9-10, pp. 488–493, 2012.

[5] P. M. Greenfield, "The changing psychology of culture from 1800 through 2000," *Psychological science*, vol. 24, no. 9, pp. 1722–1731, 2013.

[6] A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley, "Statistical laws governing fluctuations in word use from word birth to word death," *Scientific reports*, vol. 2, 2012.

[7] M. Gerlach and E. G. Altmann, "Stochastic model for the vocabulary growth in natural languages," *Physical Review X*, vol. 3, no. 2, p. 021006, 2013.

[8] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc, "Languages cool as they expand: Allometric scaling and the decreasing need for new words," *Scientific reports*, vol. 2, 2012.

[9] R. A. Bentley, A. Acerbi, P. Ormerod, and V. Lampos, "Books average previous decade of economic misery," *PloS one*, vol. 9, no. 1, p. e83147, 2014.

[10] A. Koplenig, "The impact of lacking metadata and data truncation for the measurement of cultural and linguistic change using the google ngram datasets (draft - under review)," 2014. http://hdl.handle.net/10932/00-023C-DD02-76AF-FF01-9; accessed online January 5, 2014.

[11] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951.

[12] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 1948.

[13] J. Lin, "Divergence measures based on the shannon entropy," *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 145–151, 1991.

[14] Supporting Online Materials can be found at http://www.compstorylab.org/share/papers/pechenick2015a/.

[15] D. J. de Solla Price, *Little Science, Big Science*. New York: Columbia University Press, 1963.

Chapter 3

Is language evolution grinding to a halt?: Exploring the life and death of words in English fiction

The Google Books corpus contains millions of books in a variety of languages. Due to its incredible volume and its free availability, it is a treasure trove for linguistic research. In a previous work, we found the unfiltered English data sets from both the 2009 and 2012 versions of the corpus are both heavily saturated with scientific literature, as is the 2009 version of the English Fiction data set. Fortunately, the 2012 version of English Fiction is consistent with fiction and shows promise as an indicator of the evolution of the English language as used by the general public. In this paper, we first 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. We show that this earlier method produces an artificial surge in the

death rate at the end of the observed period of time. In order to avoid this boundary effect in our own analysis of asymmetries in language dynamics, we examine the volume of word flux across various relative frequency thresholds for the 2012 English Fiction data set. We then use the contributions of the words crossing these thresholds to the Jensen-Shannon divergence between consecutive decades to resolve the major driving factors behind the flux.

3.1 Introduction

The incredible volume and free availability of the Google Books corpus [1, 2] make it an exceptional candidate for linguistic research. In a previous work [3], we explored the dynamics of the English and English Fiction data sets from both versions of the corpus, and we showed that the unfiltered 2009 and 2012 English data sets are both heavily influenced by scientific texts. In releasing the original data set, Michel et al. [1] warned that 2009 version the English Fiction contained non-fiction material, including scholarly articles about fictional works. Critically, we observed in [3] that this data set is in fact increasingly dominated over the last several decades by scientific literature with medical research language being especially prevalent. It is therefore not an appriate data set for analysis. The 2012 version of English Fiction is improved in the respect that it appears to robustly represent fiction.

In this paper, in order to avoid bias from scientific journals, we limit our analysis to the 2012 version of the English Fiction data set. Fig. 3.1 shows the total number of 1-grams for this data set between 1800 and 2000. An exponential increase in

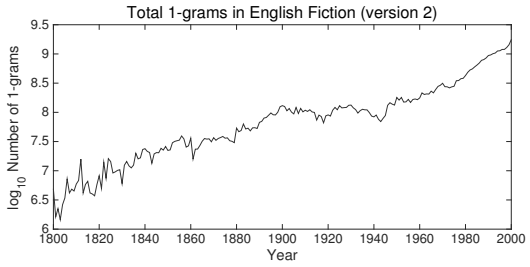

Figure 3.1: The logarithms of the total 1-gram counts for the Google Books corpus 2012 English Fiction data set. An exponential increase in volume is apparent over time with notable exceptions during wartime when the total volume decreases. (This effect is clearest during the American Civil War and both World Wars.)

volume is apparent over time with notable exceptions during major conflicts when the total volume decreases. For ease of comparison with related work, and to avoid high levels of optical character recognition (OCR) errors due to the presence of the long s—e.g., "said" being read as "faid" [3]—we omit the first two decades and focus on 1-grams between the years 1820 and 2000. We also use the terms "word" and "1-gram" interchangeably for the sake of convenience.

Many researchers have carried out broad analyses of the Google Books corpus, examining properties and dynamics of entire languages. These include analyses of Zipf's and Heaps' laws as applied to the corpus [4], the rates of verb regularization [1], rates of word "birth" and "death" and durations of cultural memory [5], as well as an observed decrease in the need for new words in several languages [6]. However, many of the studies were performed before the release of the second version and did not appear to take into account the substantial effects of scientific literature on the data sets.

The example of word "birth" and "death" rates [5] is of particular interest due to the intuitive narrative of decreased rates of introduction to vocabularies over time

and increased rates of disposal. Both observations can be explained in part by the wide availability of spell-checking software and, hence, decreased competition between alternate spellings. However, we observe that boundary effects arise from the methods employed in that paper. In particular, death rates (as defined by the authors) tend to increase as the last recorded time period is approached. (We demonstrate this in Section 3.2.) Therefore, while the conclusions drawn by the authors may be essentially correct, we must nonetheless take them with a grain of salt.

We do not, however, dispute that asymmetry exists in the changes in word use. In our earlier work [3], we observed this asymmetry in the contributions to the statistical divergences between decades with most large contributions being accounted for by words whose relative frequencies had increased. In this paper, we apply a similar information-theoretic approach to examine this effect.

We structure the remainder of the paper as follows. In Section 3.2, we critique a method from a related work [5] which examines the birth and death rates of words in a data set. In Section 3.3, we recall and confirm a similar apparent bias toward increased usage rates of words from our prevoius paper. We then measure the flux of words across various relative frequency boundaries (in both directions) in the second English Fiction data set. Furthermore, we describe the use of the largest contributions to the Jensen-Shannon divergence between successive decades from among the words crossing each boundary as signals to highlight the specific dynamics of word growth and decay over time. In Section 3.4, we display examples of these highlights and explore the factors contributing to the observed disparities between growth and decay. We offer concluding remarks in Section 3.5, summarizing the implications of our findings.

3.2 Critique of a related work

In a related paper [5], Petersen et al. examined the birth and death rates of words over time for various data sets in the first version of the Google Books corpus. They defined the birth year and death year of an individual word as the first and last year, respectively, that the given word appeared above one twentieth its median relative frequency. Excluded from considerations were words appearing in only one year and words appearing for the first time before 1700. (The latter exclusion focuses the analyis on recent words.) The rates of word birth and death, respectively, were found by normalizing the numbers of births and deaths counted by the total number of unique words in a given year.

Results typical to all data sets included decreased birth rates and increased death rates over time. As noted in the previous section, these results are not implausible. The very specific nature of the experiment—particularly the multiple temporal restrictions on the words included in the analysis, the reliance on a particular proportion of each word's median frequency, and the ignoring of all but the first and last crossings over this threshold—raise questions as to the robustness of the method. (Granted, the authors of the paper do state that the results are qualitatively similar when one tenth the median frequency is used as a threshold.)

Ignoring all but the first and last crossings, in particular, appears to cause problems. A boundary effect can arise when the death of a word is defined as the last observed occurrence of a given word above its threshold. To demonstrate this, we recreate the described analysis for the second version of English Fiction.

We note that in our analyses, the relative frequencies are coarse-grained at the level of decades. Excluded are words appearing in only one decade (rather than year) and words appearing before the 1820s (instead of 1700). Again, this more recent initial cut-off date accounts for the high frequency of OCR errors observed before 1820.

Furthermore, we compare the birth and death rates as observed recently versus historically by performing the analysis with three different endpoints imposed: the 1950s, the 1970s, and the 1990s. We present the results of the recreation in Fig. 3.2 (c.f. Fig. 2 in [5]).

The observed birth rates are qualitatively similar to those from various data sets (from the 2009 version of the corpus) in the afore-mentioned paper and display spikes in the 1890s and 1920s. The observed death rates with the 1990s boundary (light gray) are also similar, despite the lack of deaths detected during much of the 19th century. (Recall, we ignored words originating prior to 1820.) However, as the later boundary is moved to the 1970s, what was originally a stable region between the 1910s and 1940s turns into a region of gradually increasing word death. As the boundary is moved to the 1950s, the increase in death rate is no longer gradual. This demonstrates a qualitative dependence of the observations of the death rate on when the history of the corpus ends. The results of the experiment depend on when the experiment is performed.

Therefore, while this method provides a good start to analyzing asymmetries in the evolutionary dynamics of a language data set, the results of the method invite supplementation by more comprehensive methods, which we introduce presently.

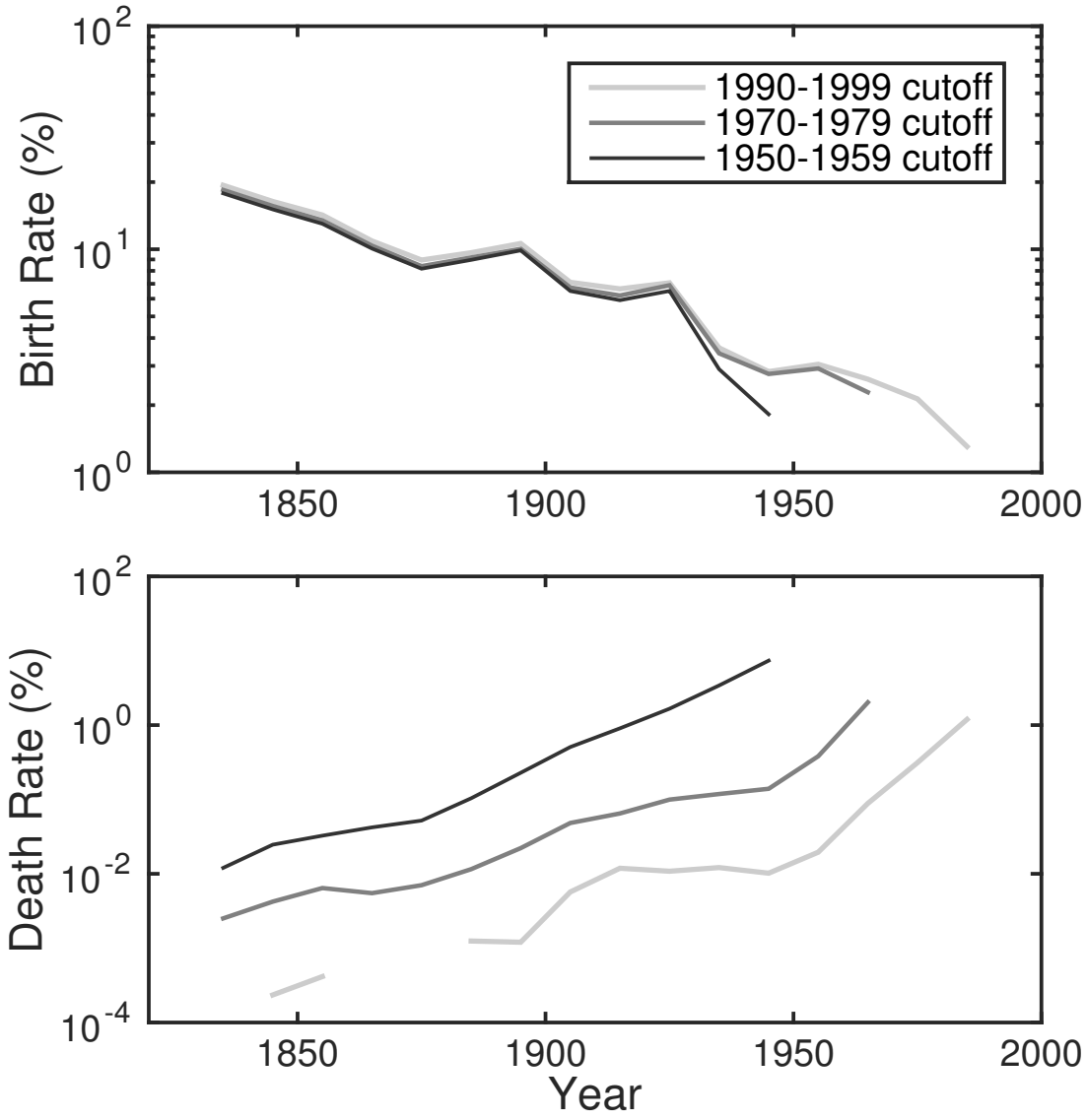

Figure 3.2: Birth and death rates, with definitions based on a related paper [5], for 2012 version of English Fiction as observed between the 1820s and three different end-of-history boundaries. The observed birth rates are qualitatively similar to those from various (2009) data sets (see Fig. 2 in the afore-mentioned paper) and display spikes in the 1890s and 1920s. The observed death rates with the 1990s boundary (light gray) are also similar, albeit with no deaths detected during much of the 19th century (a result of ignoring words originating prior to 1820). However, as the latter boundary is moved to the 1970s, what was originally a stable region between the 1910s and 1940s turns into a region of gradually increasing word death. As the boundary is moved to the 1950s, the increase in death rate is no longer gradual. This demonstrates a qualitative dependence of the observations of the death rate on when the history of the corpus ends.

3.3 Methods

We course-grain the relative frequencies in the second English Fiction data set at the level of decades—e.g., between 1820-to-1829 and 1990-to-1999—by averaging the relative frequency of each unique word in a given decade over all years in that decade. (We weight each year equally.) This allows us to conveniently calculate and sort contributions to divergence of individual 1-grams between any two time periods.

3.3.1 Statistical divergence between decades

As in our previous paper [3], we examined the dynamics of the 2012 version of English Fiction by calculating contributions to the Jensen-Shannon divergence (JSD) [7] between the distributions of 1-grams in two given decades. We then used these contributions to resolve specific and important signals in dynamics of the language. (This material, which is presented in greater detail in our previous work, is outlined in sufficient detail below.)

Given a language with 1-gram distributions P in the first decade and Q in second, the JSD between P and Q can be expressed as

$$D_{JS}(P||Q) = H(M) - \frac{1}{2}\Big(H(P) + H(Q)\Big), \tag{3.1}$$

where $M = \frac{1}{2}(P + Q)$ is a mixed distribution of the two years, and $H(P) = -\sum_i p_i \log_2 p_i$ is the Shannon entropy [8] of the original distribution. The JSD is symmetric and bounded between 0 and 1 bit. These bounds are only observed when the distributions are identical and free of overlap, respectively.

The contribution from the i^{th} word to the divergence between two decades, as derived from Eq. 3.1, is given by

$$D_{JS,i}(P||Q) = m_i \cdot \frac{1}{2}\Big(r_i \log r_i + (2 - r_i) \log(2 - r_i)\Big), \tag{3.2}$$

where $r_i = p_i/m_i$, so that contribution from an individual word is proportional to both the average frequency of the word and also depends on the ratio between the smaller and average frequencies. To elucidate the second dependency, we reframe the contribution as

$$D_{JS,i}(P||Q) = m_i C(r_i). \tag{3.3}$$

Words with larger average frequencies yield larger contribution signals as do those with smaller ratios, r_i, between the frequencies. A common 1-gram changing subtly can produce a large signal. So can an uncommon or new word given a sufficient shift from one decade to the next. $C(r_i)$, the proportion of the average frequency contributed to the signal, is concave (up) and symmetric about $r_i = 1$, where the frequency remains unchanged yielding no contribution. If a word appears or disappears between two decades (i.e., $p_i = r_i = 0$), then the contribution is maximized at precisely the average frequency of the word in question.

3.3.2 Exploring asymmetric dynamics

We observed in a previous paper [3] that most large JSD contribution signals are due to words whose relative frequencies increase over time. In this paper, we confirm and explore this effect.

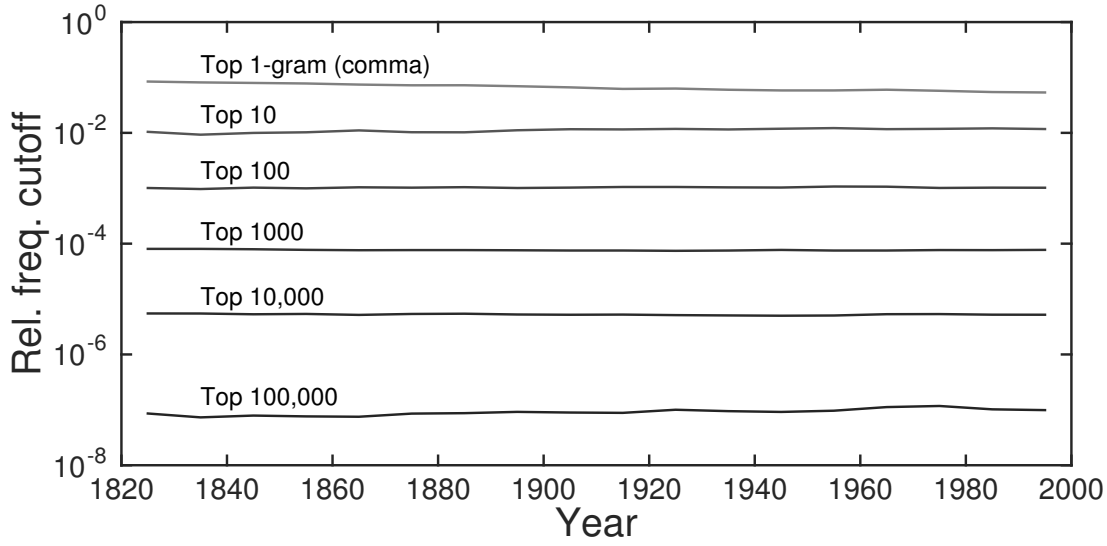

Figure 3.3: Rank threshold boundaries correspond to nearly constant relative frequency threshold boundaries over many orders of magnitude, with the exception of the top 1-gram (always a comma), which decreases in relative frequency. This demonstrates the general consistency of recording measurements related to flux across either type of boundary.

We texture our observations by examining JSD signals due to words crossing various relative frequency thresholds in either direction, as well as the total volume of word flux in either direction across these thresholds. It is both convenient and consistent to record flux over relative frequency thresholds instead of rank thresholds. To demonstrate this consistency, we observe in Fig. 3.3 that rank treshold boundaries correspond to nearly constant relative frequency thresholds, with the exception of the top 1-gram (always the comma), which decreases gradually in relative frequency. For thresholds of 10^{-5} and below, we omit signals corresponding to references to specific years, since such references would otherwise overwhelm the charts for these thresholds.

3.4 Results and Discussion

As seen in Fig. 3.4, more than half of the JSD between a typical given decade and the next is due to contributions from words increasing in relative usage frequency. The JSDs between 1820s, 1840s, and 1970s and their successive decades are the only exceptions. Moreover, when the time differential is increased to three decades, no exceptions remain. This confirms asymmetry exists between signals for words increasing and decreasing in relative use. We note relative extrema of the inter-decade JSD in the vicinity of major conflicts. Between the 1860s and successive decades, words on the rise contribute substantially to the JSD. This is consistent with words not relatively popular during wartime (specifically the American Civil War) being used more frequently in peacetime. A similar tendency holds for the JSD between the 1910s (World War I) and the 1920s. This is not as apparent in the JSD between the 1910s and the 1940s, possibly because the 1940s coincide with World War II. The absolute maximum for the single-decade curve corresponds to the divergence between the 1950s and 1960s. This suggests a strong effect from social movements. (For the 3-decade split, the absolute peak comes from the JSD between the 1940s and 1970s.)

We next consider flux across relative frequency thresholds of powers of 10 from 10^{-2} down to 10^{-6}. In Fig. 3.5, we display the volume of flux of words in both directions across relative frequency thresholds of powers of 10 from 10^{-4} down to 10^{-7}. Flux across the 10^{-2} boundary between consecutive decades is almost nonexistent during the observed period. Between the 1820s and 1830s, the semicolon falls below the threshold. Between the 1840s and 1850s, "I" rises above the boundary. Between

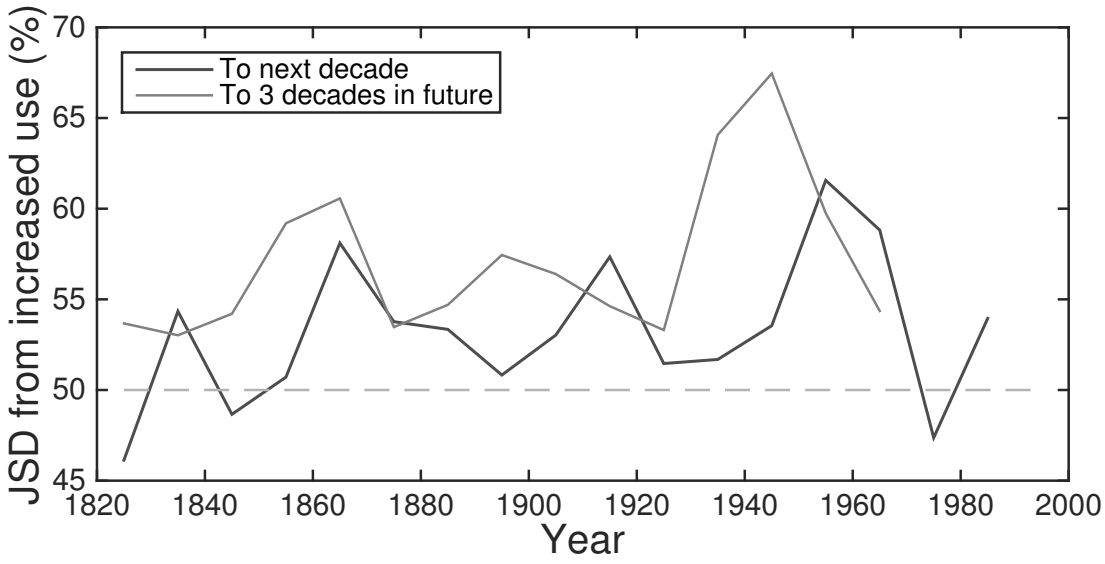

Figure 3.4: Percent of JSD in English Fiction (version 2) due to words increasing in relative frequency of use. The JSD between successive decades is nearly always more than half. The only exceptions are between the 1820s, 1840s, and 1970s, and their successive decades. When the distance between time periods is increased to 3 decades, no exceptions remain. The JSD between successive decades also shows peaks in the vicinity of major conflicts.

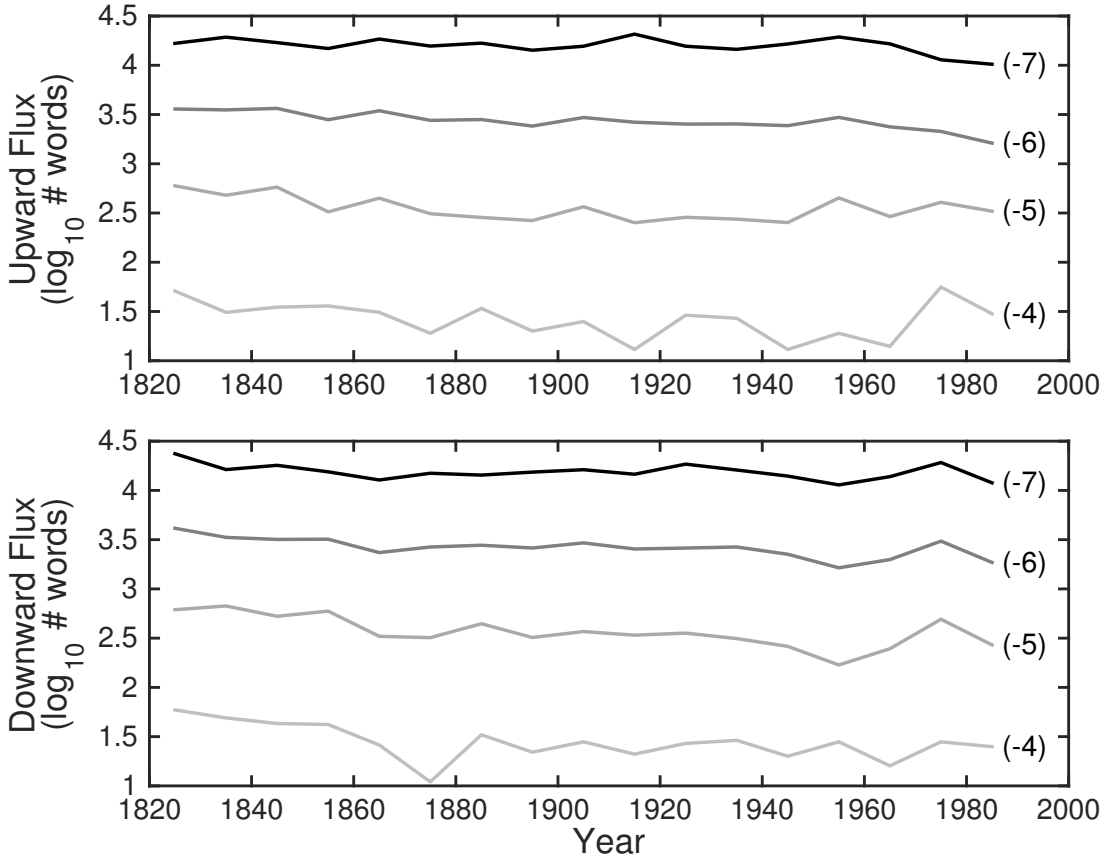

Figure 3.5: Total number of words ($\log_{10}$) crossing relative frequency thresholds of 10^{-4}, 10^{-5}, 10^{-6}, and 10^{-7} in both directions between each decade and the next decade. For each threshold, the upward and downward flux roughly cancel. For either direction of flux, there appears to be little qualitative difference between the three smallest thresholds for which the downward flux between the 1950s and the 1960s is a minimum, the downward flux increases over the next two pairs of consecutive decades, then it dips again between the 1980s and 1990s. For the highest threshold, the increase between the 1960s and 1970s and the next pair of decades is more noticeable for the upward flux, as is the decrease between the last two pairs of decades.

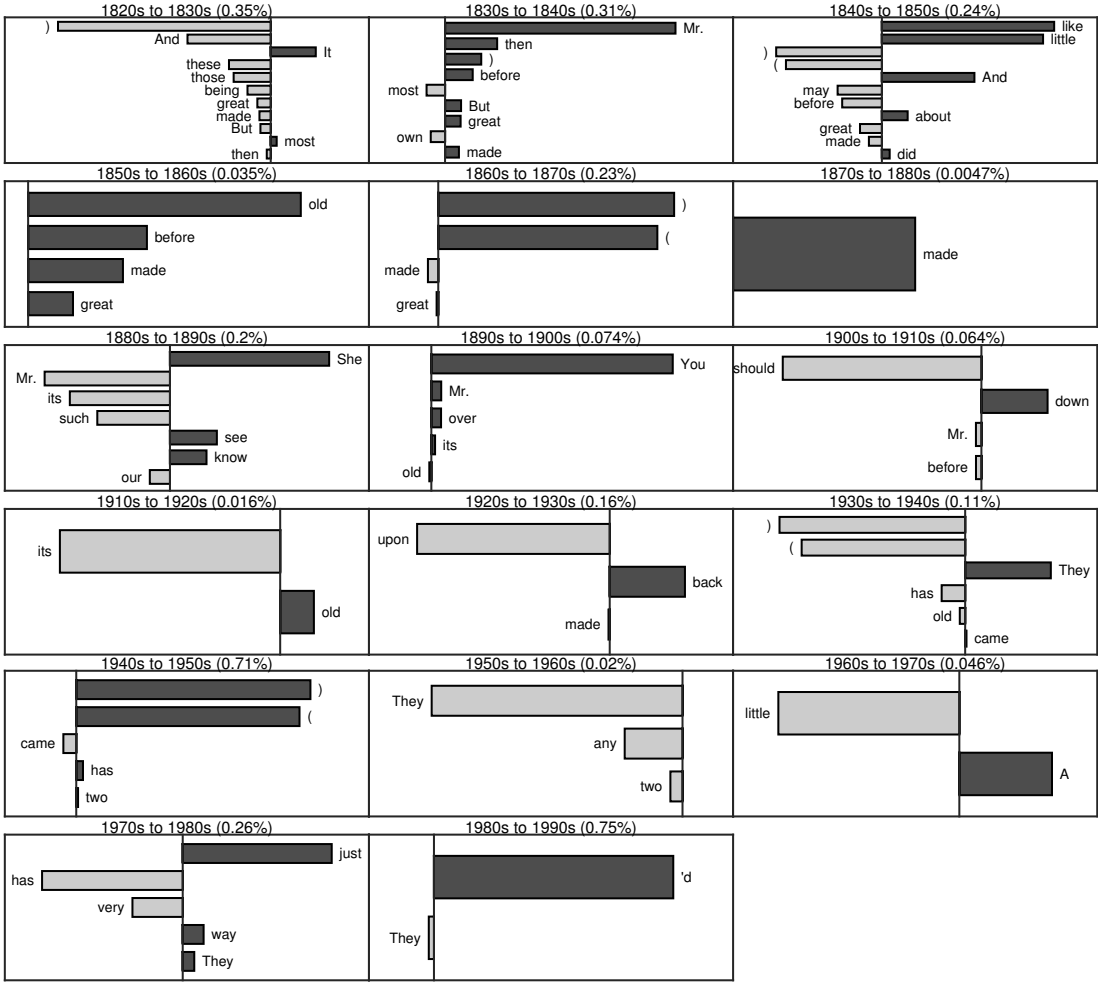

Figure 3.6: Words crossing relative frequency threshold of 10^{-3} between consecutive decades. Signals for each pair of decades are sorted and weighted by contribution to the JSD between those decades. Bars pointing to the right represent words that rose above the threshold between decades. Bars pointing left represent words that fell. In parentheses in each title is the total percent of the JSD between the given pair of decades that is accounted for by flux over the 10^{-3} threshold.

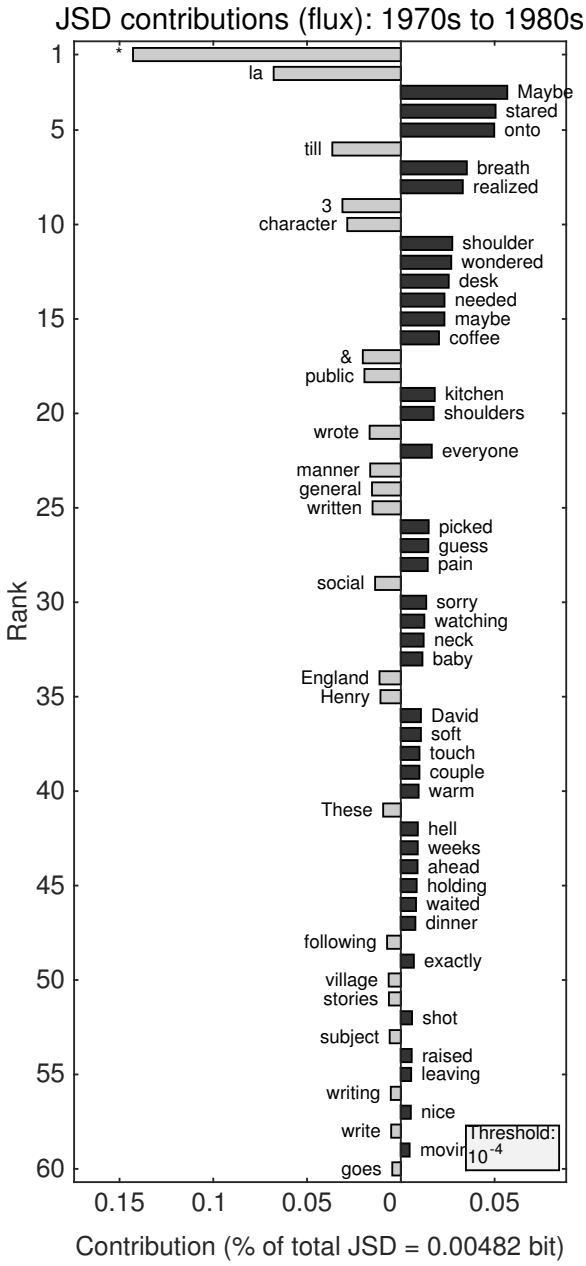

Figure 3.7: Words crossing relative frequency threshold of 10^{-4} between the given decades. Signals for each pair of decades are sorted and weighted by contribution to the JSD between those decades. Bars pointing to the right represent words that rose above the threshold between decades. Bars pointing left represent words that fell. (The first signal is the asterisk "*".)

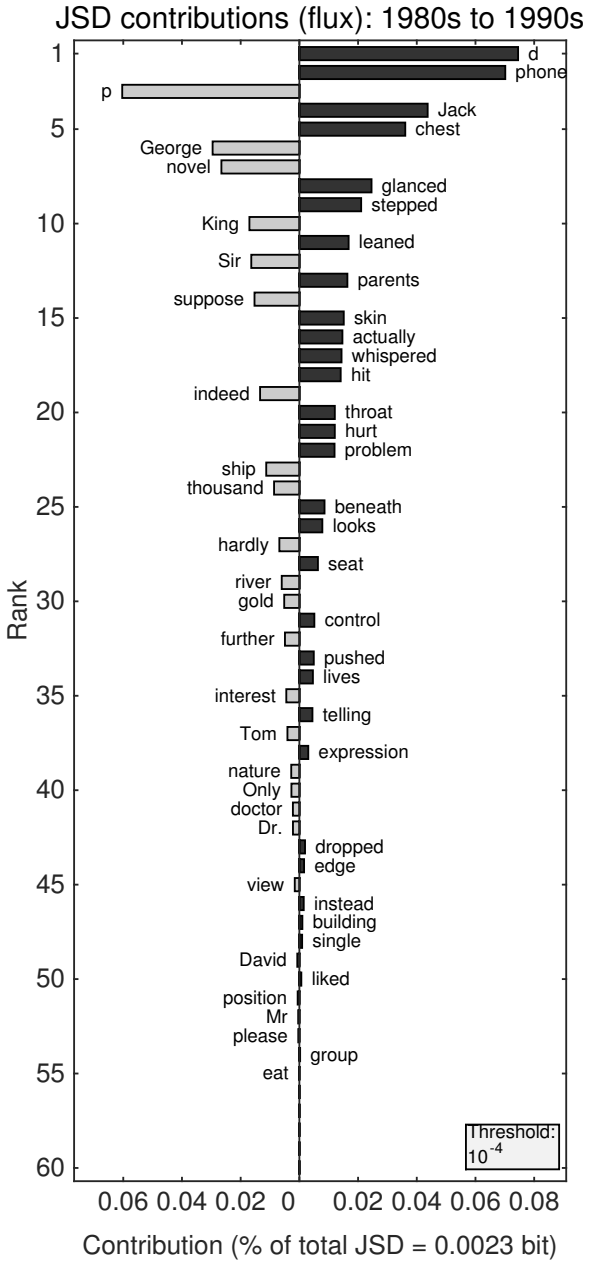

Figure 3.8: Words crossing relative frequency threshold of 10^{-4} between the given decades. Signals for each pair of decades are sorted and weighted by contribution to the JSD between those decades. Bars pointing to the right represent words that rose above the threshold between decades. Bars pointing left represent words that fell.

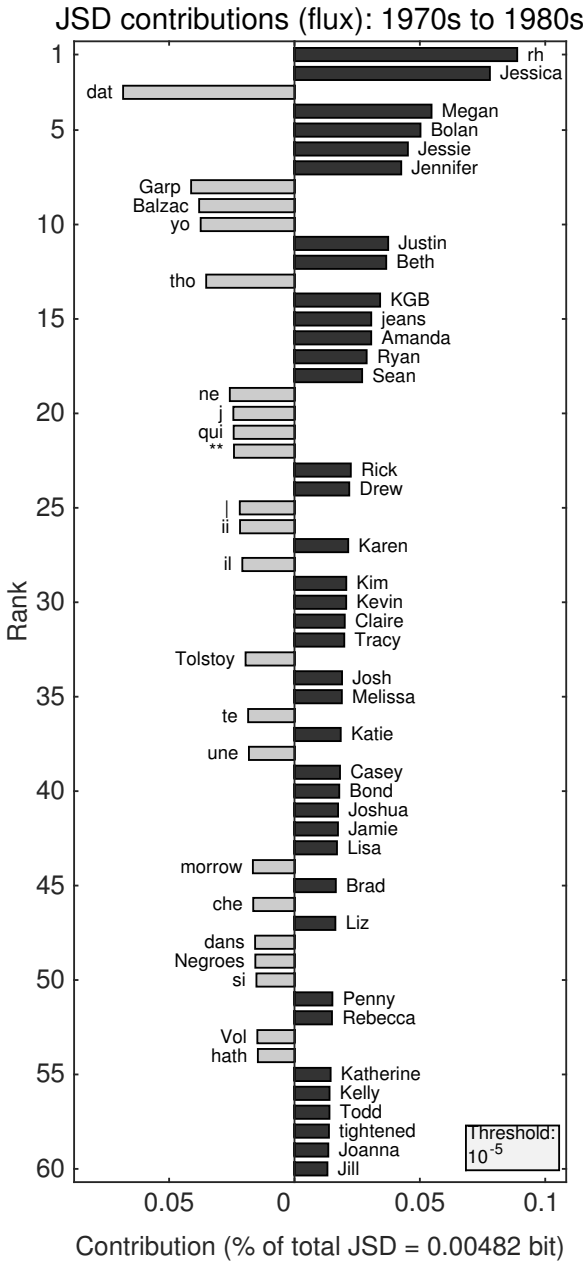

Figure 3.9: Words (not counting references to years) crossing relative frequency threshold of 10^{-5} between the given decades. Signals for each pair of decades are sorted and weighted by contribution to the JSD between those decades. Bars pointing to the right represent words that rose above the threshold between decades. Bars pointing left represent words that fell.

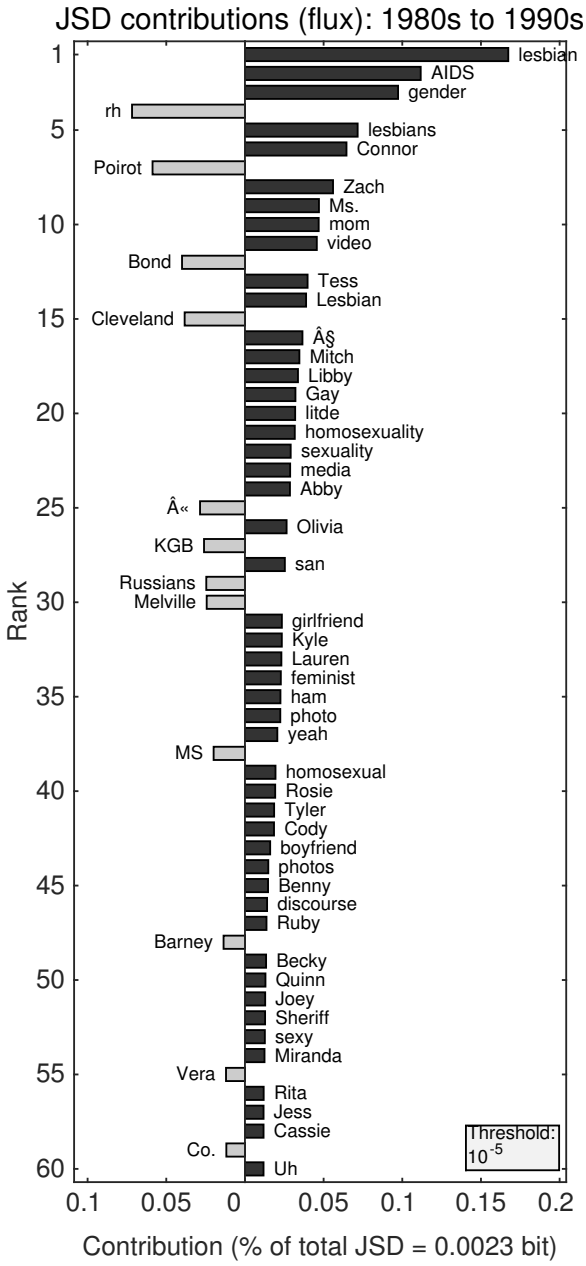

Figure 3.10: Words (not counting references to years) crossing relative frequency threshold of 10^{-5} between the given decades. Signals for each pair of decades are sorted and weighted by contribution to the JSD between those decades. Bars pointing to the right represent words that rose above the threshold between decades. Bars pointing left represent words that fell.

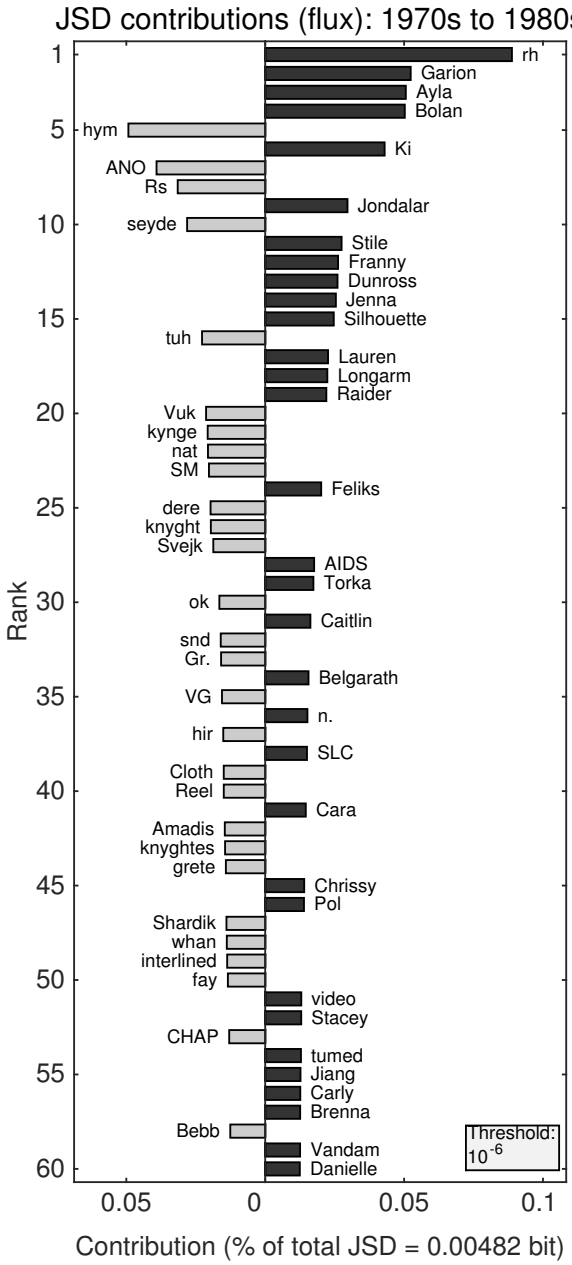

Figure 3.11: Words (not counting references to years) crossing relative frequency threshold of 10^{-6} between the given decades. Signals for each pair of decades are sorted and weighted by contribution to the JSD between those decades. Bars pointing to the right represent words that rose above the threshold between decades. Bars pointing left represent words that fell.

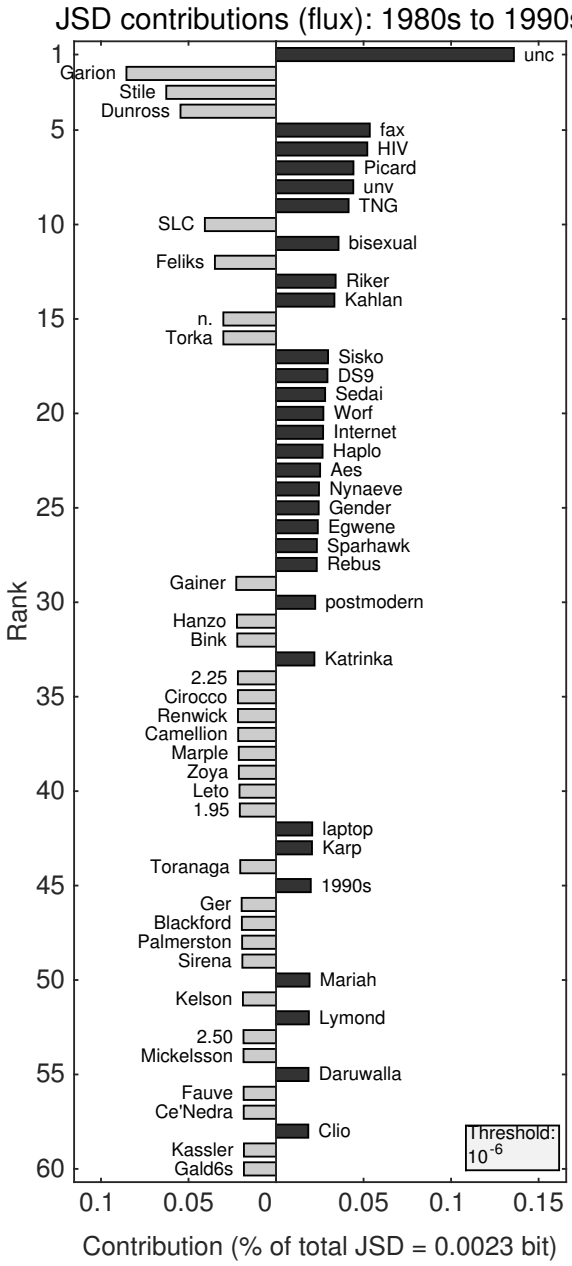

Figure 3.12: Words (not counting references to years) crossing relative frequency threshold of 10^{-6} between the given decades. Signals for each pair of decades are sorted and weighted by contribution to the JSD between those decades. Bars pointing to the right represent words that rose above the threshold between decades. Bars pointing left represent words that fell.

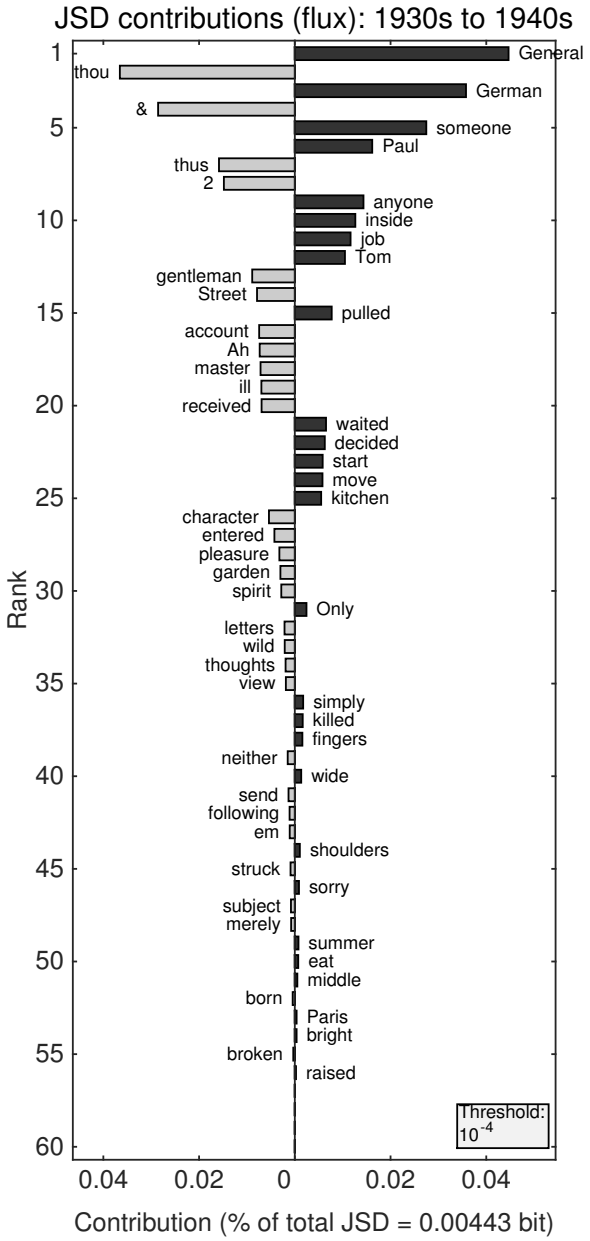

Figure 3.13: Words crossing relative frequency threshold of 10^{-4} between the given decades. Signals for each pair of decades are sorted and weighted by contribution to the JSD between those decades. Bars pointing to the right represent words that rose above the threshold between decades. Bars pointing left represent words that fell.

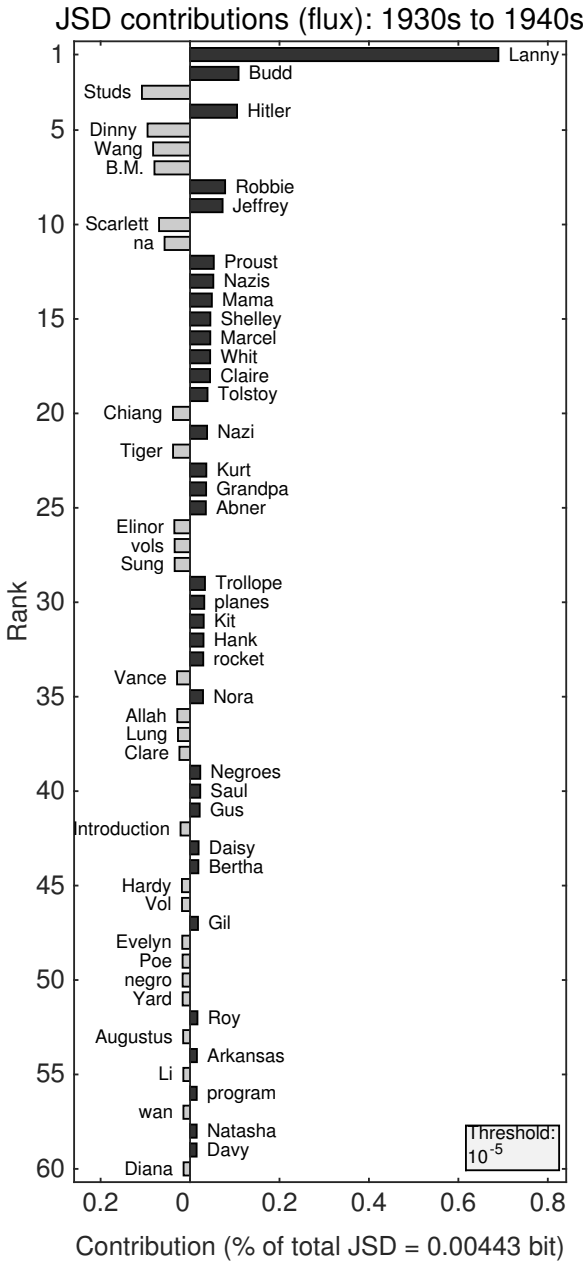

Figure 3.14: Words (not counting references to years) crossing relative frequency threshold of 10^{-5} between the given decades. Signals for each pair of decades are sorted and weighted by contribution to the JSD between those decades. Bars pointing to the right represent words that rose above the threshold between decades. Bars pointing left represent words that fell.

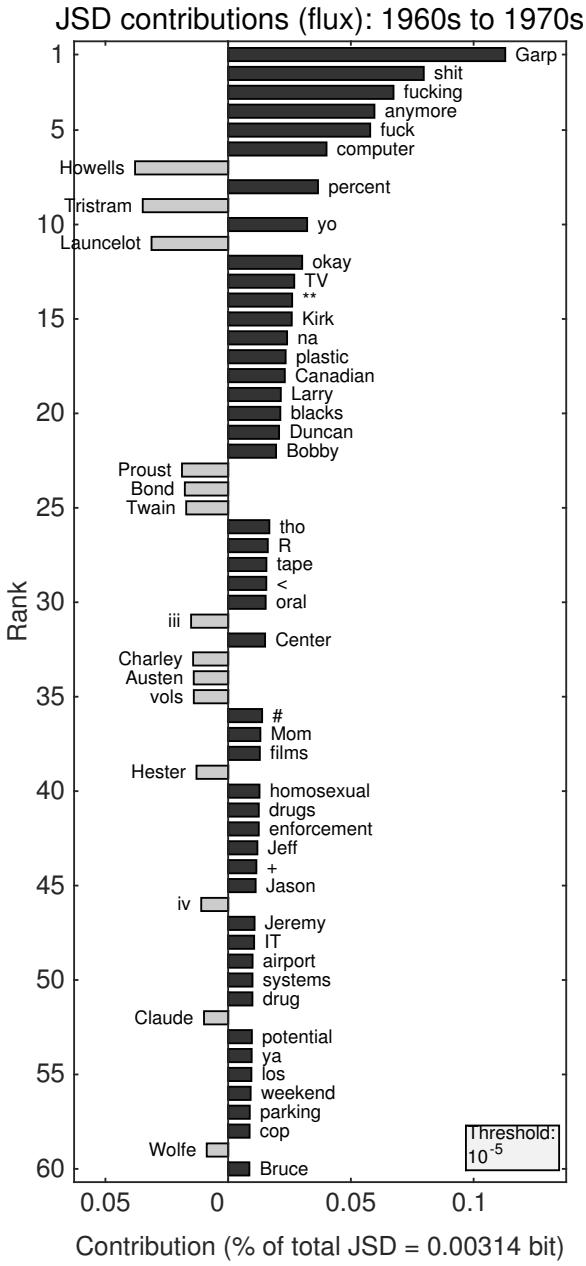

Figure 3.15: Words (not counting references to years) crossing relative frequency threshold of 10^{-5} between the given decades. Signals for each pair of decades are sorted and weighted by contribution to the JSD between those decades. Bars pointing to the right represent words that rose above the threshold between decades. Bars pointing left represent words that fell.

the 1910s and 1920s, "was" rises across. This is the entirety of the flux across 10^{-2}, which shows the regime of words above this frequency (roughly the top 10 words) is quite stable. The eleven words above threshold in the 1990s in decreasing order of frequency are: the comma ",", the period ".", "the", quotation marks, "to", "and", "of", "a", "I", "in", and "was".

The set of words with relative frequencies above 10^{-3} (roughly to the top 100 words) is also fairly stable. The flux of words across this boundary between consecutive decades is entirely captured by Fig. 3.6. Parentheses drop in (relative frequency of) use between the 1840s and 1850s and cross back over the threshold after the American Civil War (between the 1860s and 1870s). The same is true for before and after World War II (between the 1930s and 1940s and between the 1940s and 1950s, respectively). Beyond these, the flux is entirely due to proper words (not punctuation). For example, "made" fluxuates up and down over this threshold repeatedly over the course of a century. (Between the 1870s and the 1880s, "made," which sees slighty increased use, is the only word to cross the threshold. The most crossings is 12, which occurs between the first two decades.) Also, "great" struggled over the first 5 decades and eventually failed to remain great by this measure. "Mr." fluctuated across the threshold between the 1830s and 1910s. More recently (since the 1930s), "They" has been making its paces up and down across the threshold.

For each threshold between 10^{-4} and 10^{-7}, the upward and downward flux roughly cancel, which is consistent with Fig. 3.3. For both upward and downward flux, there appears to be little qualitative difference between the three smallest thresholds. For these thresholds, the downward flux between the 1950s and the 1960s is a minimum, the downward flux increases over the next two pairs of consecutive decades, then

it dips again between the 1980s and 1990s. For the highest threshold, the increase between the 1960s and 1970s and the next pair of decades is more noticeable for the upward flux, as is the decrease between the last two pairs of decades.

In the experiment recreated in Fig. 3.2, the word birth rate initially exceeds the death rate by three orders of magnitude, and this gap declines gradually over the next two centuries. However, with respect to words fluxuating across relative frequency thresholds in opposite directions, we see no strong evidence of asymmetry during any long period of time. With respect to total contributions to the JSD between consecutive decades, there is typically some bias toward words with increased relative use as seen in Fig. 3.4, but the difference need never be described in orders of magnitude.

To address the fluxuations during the last couple of decades, we begin by displaying in Fig. 3.7 the top 60 flux words between the 1970s and the 1980s sorted by contributions to the JSD between those decades. Note that this pair of decades corresponds to both a dip (below 50%) in the proportion of rising word contributions to the JSD and to an increase in the volume of downward flux (as well as upward flux for high thresholds). In Fig. 3.8, we show all 55 flux words between the 1980s and the 1990s.

Between each pair of decades, we see reduced relative use of particulary British words, including "England" between the first two decades and "King," "George," and "Sir" between the latter two. We also see reduced use of more formal-sounding words, such as "character," "manner," and "general" between the first two decades and "suppose," "indeed," and "hardly" between the latter two. Increasing are physical and emotional words. Those between the first two decades include "stared," "breath," "realized," "shoulder" and "shoulders," "coffee," "guess," "pain," and "sorry." Be-

tween the latter two, we see "chest," "skin," "whispered," "hit," "throat," "hurt," "control," and "lives." Also included are "phone" and "parents."

In Figs. 3.9 and 3.10, we display the top 60 flux words, not counting references to years, across the 10^{-5} threshold between the same decades. Many of the words declining below the threshold between the 1970s and 1980s are unusual spellings such as "tho," proper names like "Balzac," or words from non-English languages like "une." Also included is the word "Negroes." Increasing across this threshold between the first two decades are a plethora of mostly female proper names, with "Jessica" and "Megan" leading. Also seen are "KGB" and "jeans." ("KGB" decreases in the 1990s, as does "Russians.") Increasing between the 1980s and 1990s are a few proper names; however, most of the signals here are social and sexual in nature. These include "lesbian" and "lesbians," "AIDS," and "gender" in the top positions. Also included are both "homosexuality" and the more general "sexuality." We also see "girlfriend," "boyfriend," "feminist," and "sexy."

For contrast and amusement, we show in Fig. 3.12 the flux across a threshold of 10^{-6} between the 1980s and 1990s (again, not counting years). In particular, while increases in "HIV" and "bisexual" make the list (similarly to many signals in Fig. 3.10), as do "fax," "laptop," and "Internet," a great swath of the signals are accounted for by one franchise. Note increases in "Picard," "TNG," "Sisko," and "DS9." These latter signals should serve as a reminder that the word distributions in the corpus, even for fiction, do not always resemble the contents of normal conversations (at least not for the general population). However, we do observe signals arising at this threshold from factors external to the imaginings of specific authors. It would therefore be premature to dismiss the contributions at this threshold because of an apparent over-

abundance of "Star Trek." In fact, since "The Next Generation" and "Deep Space 9" aired precisely during these two decades, an abundance of "Star Trek" novels in the English Fiction data set is actually quite encouraging, because these novels do exist, are available in English, and are (clearly) fiction.

For consistency, we also include the flux (omitting years) across this threshold between the 1970s and 1980s in Fig. 3.11. While not particularly topical, we do see "AIDS" increase above this threshold a decade prior to its increase over 10^{-5} as seen in Fig. 3.10.

The texture of the signals changes as we dial down the frequency threshold. We typically find that thresholds of 10^{-4} and above produce signals with little to no noise. This is not surprising since this relative frequency roughly corresponds to rank threshold for the 1000 most common words (see Fig. 3.3) in the data set. Using a threshold of 10^{-5} (fewer than 10,000 words fall above this frequency in any given decade), we see some noise (mostly in the form of familiar names), but still observe many valuable signals. Only when the threshold is reduced to 10^{-6} does the overall texture of the signals become questionable as a result of a variety of proper nouns far less familiar than those observed with the previous threshold. However, at this threshold, we also observe several early signals of real social importance.

Curiously, between the 1930s and 1940s the volume of flux across each threshold is not atypical (see Fig. 3.5). Moreover, the asymmetry between the JSD contributions between those decades is very low. Yet it is obvious that we should expect signals of historical significance between these two decades. In Figs. 3.13 and 3.14, we see words crossing the 10^{-4} and 10^{-5} thresholds, respectively (with references to years omitted in Fig 3.14). For the higher threshold, only 56 words cross. The most noticeable such

words that are more commonly used in the 1940s are "General" and "German." Also, "killed" appears in this list. Words used less frequently include "pleasure," "garden," and "spirit." For the lower threshold, we see the signals from prolific authors as in our previous paper [3], particularly Upton Sinclair's character, Lanny Budd. We also see more Nazis ("Nazi" and "Nazis"). In fact, the Lanny Budd series is a war story, meaning even the author-specific works are relevant to the era (much like the "Star Trek" example earlier).

Last, we include one of the more colorful examples. In Fig. 3.15, we show signals (not including years) for words crossing the 10^{-5} threshold between the 1960s and 1970s. Profanity dominates. We see more references to *The World According to Garp* ("Garp") and "Star Trek," again ("Kirk" this time). We also see more "computer," "TV," and "plastic." Signals also appear for "blacks" and "homosexual," for drugs ("drug" and "drugs"), and (plausibly) for the War on Drugs ("enforcement" and "cop").

See the Supporting Online Materials [9] for figures representing flux across relative frequency thresholds of 10^{-4}, 10^{-5}, and 10^{-6} between consecutive decades over the entire period analyzed (the 1820s to the 1990s).

3.5 Concluding remarks

We recall from a related work [6] and from our own work [3] (Fig. 7d) that the rate of change of given language tends to slow down over time. This applies to the 2012 English Fiction data set and is not contested by us in the present paper. In the critiqued paper [5], it was suggested that the birth and death rates of words can

be calculated in an intuitive, albeit very specific manner. This experiment produces birth rates that begin vastly higher than death rates with both rates converging over time to around 1%. However, we have seen that these rates converge to roughly the same values at the end of the available history, regardless of when that is—i.e., the experiment depends on when you perform it, and recent results always appear qualitatively similar.

Beyond this boundary issue, we find another cause for concern. When the increased usage bias in the JSD contributions and the overall and directed volumes of flux are taken into account, we do not observe even the initial orders-of-magnitude gap between so-called birth and death rates. Rather, the JSD bias toward increased relative use of words is within one order of magnitude, and the flux across thresholds is typically balanced.

In fact, this latter point appears to be a fundamental facet of this data set. As we see in Fig. 3.3, the number of words above each threshold is roughly constant. This stability of the rank-frequency relation compels the observed balancing act (and is consistent with a stable Zipf law distribution [10]). Previously [3] (Fig. 5d), we have seen the divergence between a given year—e.g., 1880—and a target year tends to increase gradually with the time difference. This is not true when, for example, the target year—e.g., 1940—falls during a major war, in which case we see a spike in divergence. However, as the target year exits this period—e.g. enters the 1950s—the spike settles back into the original gradual growth pattern. It is plausible based on these earlier observations and the observations in this paper that the distribution of the language is self-stabilizing: the overall shape of the distribution does not appear

to change drastically with time or with the total volume of the data set. As old words fall out of favor, new words inevitably appear to fill in the gaps.

Furthermore, despite the fact that the divergence between consecutive years has been observed to decay over time, we find no shortage of novel word introductions during the most recent decades (which have the lowest decade-to-decade JSDs). This apparent dissonance clearly invites further investigation.

Finally, while extremely specific fiction can be of great interest—whether it be in the form of war novels or volumes from the "Star Trek" franchise—vocabulary from these works is more easily studied when placed in proper context. Dialing down the relative frequency threshold across several orders of magnitude helps to capture this distinction. However, further experimentation is invited, since an automatic means of separating specific signals from the more general signals (e.g., "Star Trek" from social movements) could allow both a more intuitive grasp of the linguistic dynamics and might, ideally, allow investigators to hypothesize causal relationships between exogenous and endogenous drivers of the language.

References

[1] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

[2] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the google books ngram corpus," in *Proceedings of the ACL 2012 System Demonstrations*, pp. 169–174, Association for Computational Linguistics, 2012.

[3] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, "Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution," *arXiv preprint arXiv:1501.00960*, 2015.

[4] M. Gerlach and E. G. Altmann, "Stochastic model for the vocabulary growth in natural languages," *Physical Review X*, vol. 3, no. 2, p. 021006, 2013.

[5] A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley, "Statistical laws governing fluctuations in word use from word birth to word death," *Scientific reports*, vol. 2, 2012.

[6] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc, "Languages cool as they expand: Allometric scaling and the decreasing need for new words," *Scientific reports*, vol. 2, 2012.

[7] J. Lin, "Divergence measures based on the shannon entropy," *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 145–151, 1991.

[8] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 1948.

[9] Supporting Online Materials can be found at http://www.compstorylab.org/share/papers/pechenick2015b/.

[10] G. K. Zipf, "Human behavior and the principle of least effort.," 1949.

Chapter 4

Main plots and subplots in Google Books: Exploring the effects of conflict on English fiction

The Google Books corpus is a valuable resource for linguistic research. In two earlier works, we used an information-theoretic approach to explore the dynamics of both versions of the English and English Fiction data sets. The second version of English Fiction proved to be the only one of these that did not display an obvious bias toward scientific literature. For this data set, the largest contributions to the Jensen-Shannon divergence between decades yielded a variety of fascinating words for additional study, as did the largest signals for words crossing various relative frequency thresholds. Critically, no specific topics were chosen beforehand, so these

word selections are principled. In this work, we explore the distinction between words with broad importance and interesting words with narrower ramifications by contrasting the lists of words most highly correlated with specific examples of exogenous signals and endogenous signals. We also examine the effects of war on the second version of English Fiction.

4.1 Introduction

Previously [1, 2], we used an information-theoretic approach to explore the dynamics of both versions of the English and English Fiction data sets in the Google Books corpus [3, 4]. The 2012 version of English Fiction proved to be the only one of these that did not display an obvious bias toward scientific literature, particularly in recent decades.

The diversity [5] of a given distribution, P, is the number of uniformly distributed species that would display the same amount entropy as the given distribution. This can be thought of as the number of critical members of the distribution in question. For Shannon entropy [6], $H(P) = -\sum_i p_i \ln p_i$, the Shannon diversity is $D(P) = e^{H(P)}$.

In Fig. 4.1, we compare the number of unique words present in the 2012 English Fiction data set between 1840 and 2000 to the Shannon diversity of the word distributions. (Note: We use "word" and "1-gram" interchangeably.) As with the total volume of words (c.f. Fig. 1 in [2]), we observe an exponential increase in unique words over time with notable exceptions during wartime when the total volume decreases. The word diversity, in contrast, increases gradually over time (remaining on the order

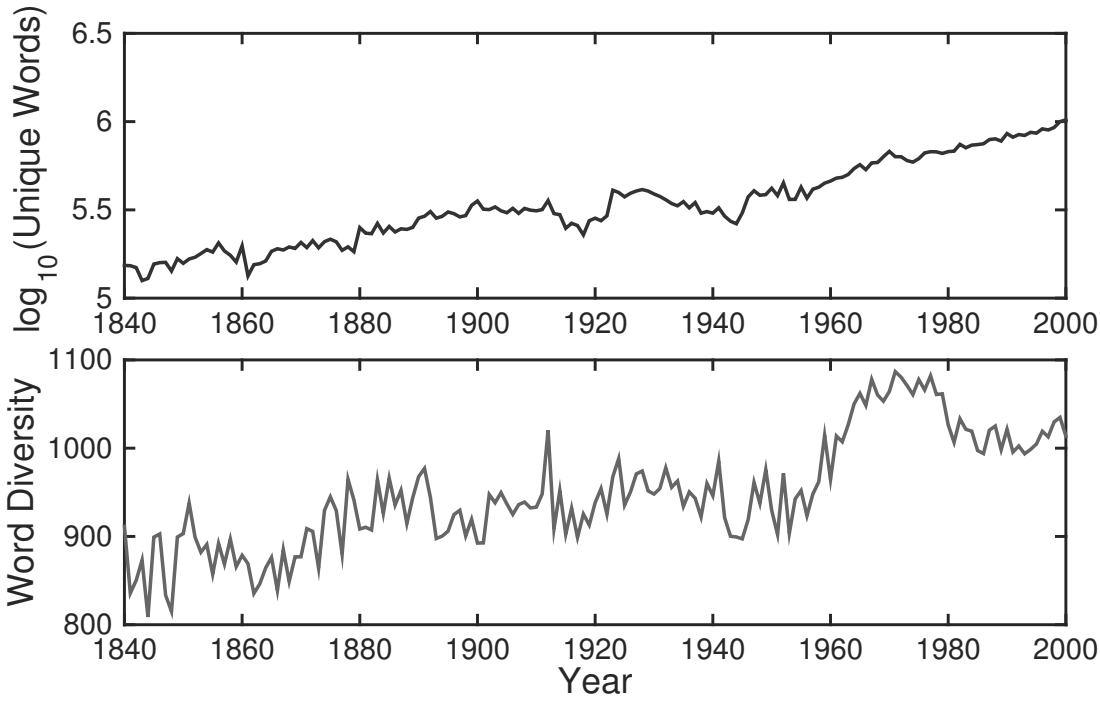

Figure 4.1: (Top) The logarithms of the unique 1-gram counts for the Google Books corpus 2012 English Fiction data set. (Bottom) The Shannon diversity of the 1-gram distribution. An exponential increase in unique 1-grams is apparent over time with notable exceptions during wartime when the total volume decreases. The word diversity increases gradually over time except during wartime when it stagnates and during the 1960s and 1970s when it is noticeably higher than usual.

of 1000 throughout) except during wartime when it stagnates and during the 1960s and 1970s when it is noticeably higher than usual.

In our afore-mentioned works [1, 2], we found both "Hitler" and "Lanny" to be important signals for the 1940s. In fact, Lanny Budd, the main character in 11 Upton Sinclair novels published between 1940 and 1953, exceeds Hitler by this measure. In Fig. 4.2, we show the relative frequencies of "Lanny" and "Hitler" in English fiction between 1930 and 2000. While Lanny dominates during the 1940s, Hitler has a steadily increasing presence in the 1930s and a lasting presence after the war. Since

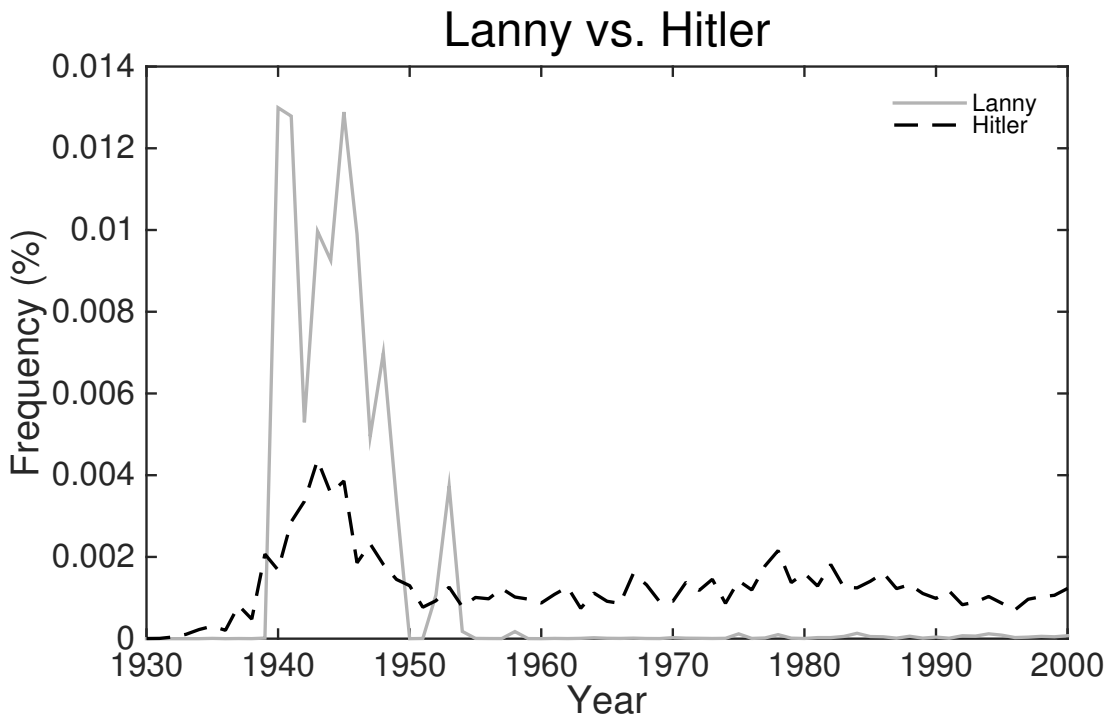

Figure 4.2: Relative frequencies of "Lanny" and "Hitler" in English fiction between 1930 and 2000. While Lanny dominates during the 1940s, Hitler has a steadily increasing presence in the 1930s and a lasting presence after the war.

the dynamics of Google Books data sets, especially English Fiction, are often driven by the contributions of prolific authors, it is necessary to distinguish words with broad and continuing importance from interesting words with narrower ramifications. In this paper, we use words discovered objectively in our previous work to explore the distinction between exogenous and endogenous signals, especially during times of conflict.

We structure the remainder of the paper as follows. In Section 4.2, we discuss the use of statisical correlations between words chosen from an objective set and relatively frequent words in the English Fiction data set to distinguish between exogenous and endogenous factors during interesting time periods. In Section 4.3, we show tables

of highest-ranked correlations for our selected words and qualitatively contrast the rankings for broadly important words and narrowly important for each considered time period. We offer concluding remarks in Section 4.4, summarizing the implications of our findings.

4.2 Methods

For the 2012 version of English Fiction, the largest contributions [1] to the Jensen-Shannon divergence between decades yielded a variety of fascinating words for additional study, as did the largest signals for words crossing various relative frequency thresholds [2]. Critically, no specific topics were chosen beforehand, which allows us to select interesting words for further analysis in a principled manner.

For each word chosen, we calculate the Spearman correlations over relevant time periods between the relative frequency time series for each given word and every word in the English Fiction data set whose relative frequency exceeds 10^{-6} at some point during that period. We select an experiment-wide significance value, $\bar{\alpha}$, for each word chosen by dividing $\alpha = 0.05$ by the number of comparisons. A Spearman correlation coefficient (ρ) very close to 1 implies a strong tendency for the frequencies of the two words being compared to rise together and fall together. (While we could also calculate Pearson correlations, which are associated with linear relationships between variables, time series in the chosen data set can be spiky, and the time series being compared often differ in orders of magnitude. Therefore, Pearson correlations may be unnecessarily restrictive.) We choose the threshold of 10^{-6} as a balance between

precision and recall. In [2] we found the signals from words crossing this threshold to be somewhat noisy, but not entirely devoid of interesting content.

As a sanity check for the method, before delving into specific periods, we list the 20 highest ranked correlations with the frequency series for "war" between 1840 and 2000 in Table 4.1. We also examine the correlations for "war" during the period from 1905 to 1925 (Table 4.2) and the period from 1930 to 1950 (Table 4.3).

To explore the distinction between exogenous and endogenous factors in the data set during World War II, we select the time series for "Lanny" and "Hitler" between 1930 and 1960 (Table 4.4). For the period between 1960 and 1980, which sees increased diversity according to Fig. 4.1, we select "Nixon" and "Spock" (Table 4.5). Again, we have a real-life individual and a fictional character. Both represent important JSD signals between the 1960s and 1970s [1]. For the period between 1980 and 2000, we contrast the correlations to "lesbian" and "Picard" (Table 4.6). The former is the most important signal crossing a threshold of 10^{-5}. The latter, along with other "Star Trek" characters, is an important signal crossing a threshold of 10^{-6} [2].

4.3 Results and Discussion

For "war" between 1840 and 2000, there are thousands of significant correlations, 20 of which we list in Table 4.1. However, the highest Spearman correlation coefficient observed for "war" across this period of time is 0.712, which is well below 1. This is not surprising, since different wars have different features, and this period contains multiple wars. We note that the highest ranked correlations suggest 20th century conflicts ("Germans" and "Allies" ranked highest), particularly World War

Table 4.1: Highest ranked Spearman correlations for "war" between 1840 and 2000. Required 142342 comparisons, $\bar{\alpha} = 3.513 \times 10^{-7}$, 6686 significant correlations. The words listed are fairly general to 20th century wars (Germans and Allies) but appear somewhat focused on World War I (especially entries 16 and 17). The highest correlation coefficient, 0.712, is well below 1 due to the large time period sampled.

war	Correlation coef. (ρ)	p-value
1) Germans	0.712	3.268×10^{-26}
2) Allies	0.687	7.86×10^{-24}
3) AMERICA	0.686	9.807×10^{-24}
4) democracy	0.672	1.604×10^{-22}
5) detachment	0.665	7.016×10^{-22}
6) Armistice	0.664	7.797×10^{-22}
7) G.H.Q.	0.662	1.145×10^{-21}
8) STATES	0.662	1.202×10^{-21}
9) cinemas	0.661	1.425×10^{-21}
10) five	0.661	1.491×10^{-21}
11) Boches	0.658	2.245×10^{-21}
12) bombardment	0.656	3.35×10^{-21}
13) telephoning	0.656	3.362×10^{-21}
14) unimportant	0.656	3.525×10^{-21}
15) airmen	0.654	5.631×10^{-21}
16) trenches	0.653	6.617×10^{-21}
17) 1918	0.652	7.028×10^{-21}
18) N.C.O.	0.651	8.873×10^{-21}
19) UNITED	0.651	9.157×10^{-21}
20) aerodrome	0.651	9.485×10^{-21}

Table 4.2: Spearman correlations for "war" between 1905 and 1925. Required 63741 comparisons, $\bar{\alpha} = 7.844 \times 10^{-7}$, 34 significant correlations. These words are unsurprisingly focused on World War I. "Kaiser" is ranked 33rd. "Tommies" at rank 12 was WWI slang for British soldiers. Several correlation coefficients are at least 0.9. The lowest is 0.856.

war	Correlation coef. (ρ)	p-value
1) Belgium	0.926	1.779×10^{-9}
2) nearby	0.917	5.157×10^{-9}
3) onto	0.9	2.793×10^{-8}
4) protests	0.896	3.953×10^{-8}
5) aircraft	0.89	6.518×10^{-8}
6) Germans	0.887	8.455×10^{-8}
7) barbed	0.886	9.375×10^{-8}
8) shoving	0.886	9.375×10^{-8}
9) focused	0.884	1.038×10^{-7}
10) munitions	0.884	1.038×10^{-7}
11) collided	0.883	1.149×10^{-7}
12) Tommies	0.883	1.149×10^{-7}
13) burlap	0.882	1.269×10^{-7}
14) Belgian	0.881	1.4×10^{-7}
15) bomb	0.878	1.7×10^{-7}
16) handicap	0.875	2.055×10^{-7}
17) sandbags	0.875	2.055×10^{-7}
18) crook	0.874	2.255×10^{-7}
19) 1916	0.873	2.473×10^{-7}
20) Irvin	0.87	2.965×10^{-7}
21) shoved	0.87	2.965×10^{-7}
22) dominated	0.869	3.241×10^{-7}
23) garbed	0.865	4.213×10^{-7}
24) wireless	0.865	4.213×10^{-7}
25) bulge	0.864	4.59×10^{-7}
26) Joffre	0.863	4.634×10^{-7}
27) defenseless	0.862	4.996×10^{-7}
28) stiffening	0.861	5.433×10^{-7}
29) atrocities	0.86	5.903×10^{-7}
30) War	0.858	6.41×10^{-7}
31) Boches	0.856	7.415×10^{-7}
32) fake	0.856	7.538×10^{-7}
33) Kaiser	0.856	7.538×10^{-7}
34) munition	0.856	7.538×10^{-7}

Table 4.3: Highest ranked Spearman correlations for "war" between 1930 and 1950. Required 63293 comparisons, $\bar{\alpha} = 7.900 \times 10^{-7}$, 63 significant correlations. These words are focused on World War II. Several correlation coefficients are at least 0.9. "Nazis," "Hitler," "Fascist," and "Mussolini" are present.

war	Correlation coef. (ρ)	p-value
1) planes	0.927	1.512×10^{-9}
2) quarreling	0.917	5.157×10^{-9}
3) plane	0.909	1.171×10^{-8}
4) A.R.P.	0.902	2.353×10^{-8}
5) Nazis	0.901	2.479×10^{-8}
6) operator	0.896	3.953×10^{-8}
7) Maginot	0.895	4.187×10^{-8}
8) reported	0.891	6.154×10^{-8}
9) doorknob	0.89	6.85×10^{-8}
10) tonight	0.89	6.85×10^{-8}
11) Hitler	0.888	7.615×10^{-8}
12) unload	0.888	7.615×10^{-8}
13) gun	0.886	9.375×10^{-8}
14) Zaharoff	0.885	9.651×10^{-8}
15) appraisal	0.884	1.038×10^{-7}
16) Fascist	0.884	1.038×10^{-7}
17) Italians	0.882	1.269×10^{-7}
18) Mussolini	0.882	1.269×10^{-7}
19) lined	0.881	1.4×10^{-7}
20) Freddi	0.88	1.436×10^{-7}

Table 4.4: Highest ranked Spearman correlations for "Hitler" and "Lanny" between 1930 and 1960. Required 72382 comparisons, $\bar{\alpha} = 6.908 \times 10^{-7}$. 105 significant correlations for "Hitler," 52 significant correlations for "Lanny."

Hitler	ρ	p-value	**Lanny**	ρ	p-value
1) Nazi	0.954	1.024×10^{-16}	1) Detaze	0.873	1.498×10^{-10}
2) Nazis	0.939	5.474×10^{-15}	2) Lannv	0.868	2.68×10^{-10}
3) Maginot	0.927	6.663×10^{-14}	3) ANY	0.865	3.392×10^{-10}
4) Fuhrer	0.92	2.508×10^{-13}	4) Nuts	0.864	3.681×10^{-10}
5) Gestapo	0.915	5.489×10^{-13}	5) Stubendorf	0.862	4.7×10^{-10}
6) Fiihrer	0.91	1.213×10^{-12}	6) Maginot	0.854	1.028×10^{-9}
7) bombing	0.91	1.292×10^{-12}	7) Blackless	0.853	1.133×10^{-9}
8) Daladier	0.899	6.347×10^{-12}	8) Budd	0.848	1.736×10^{-9}
9) Berchtesgaden	0.894	1.239×10^{-11}	9) sputtered	0.834	5.676×10^{-9}
10) bombs	0.891	1.825×10^{-11}	10) planes	0.831	7.122×10^{-9}
11) bombers	0.888	2.727×10^{-11}	11) Chattersworth	0.827	9.473×10^{-9}
12) Ribbentrop	0.885	3.716×10^{-11}	12) Hitler	0.827	9.781×10^{-9}
13) Germans	0.881	6.454×10^{-11}	13) Daladier	0.826	1.068×10^{-8}
14) Goebbels	0.873	1.52×10^{-10}	14) A.R.P.	0.825	1.145×10^{-8}
15) planes	0.869	2.22×10^{-10}	15) fix	0.825	1.16×10^{-8}
16) bombed	0.869	2.414×10^{-10}	16) gray	0.823	1.353×10^{-8}
17) Seyss	0.862	4.767×10^{-10}	17) duquesa	0.817	2.0×10^{-8}
18) colored	0.856	7.975×10^{-10}	18) war	0.817	2.027×10^{-8}
19) outdoors	0.854	9.997×10^{-10}	19) Nazis	0.811	3.129×10^{-8}
20) war	0.852	1.203×10^{-9}	20) Nazi	0.806	4.32×10^{-8}
21) Axis	0.851	1.309×10^{-9}	21) traveled	0.804	5.235×10^{-8}
22) refugees	0.848	1.729×10^{-9}	22) 1940	0.802	5.834×10^{-8}
23) 1940	0.847	1.924×10^{-9}	23) scowled	0.801	6.24×10^{-8}
24) Italians	0.846	2.065×10^{-9}	24) Rick	0.793	1.062×10^{-7}
25) A.R.P.	0.846	2.108×10^{-9}	25) Reich	0.792	1.138×10^{-7}
26) storeroom	0.842	2.922×10^{-9}	26) Berchtesgaden	0.79	1.231×10^{-7}
27) 1939	0.84	3.464×10^{-9}	27) color	0.79	1.276×10^{-7}
28) color	0.839	3.705×10^{-9}	28) Ribbentrop	0.79	1.292×10^{-7}
29) traveled	0.834	5.51×10^{-9}	29) outdoors	0.787	1.558×10^{-7}
30) Nazism	0.833	6.238×10^{-9}	30) Freddi	0.784	1.792×10^{-7}
31) snorted	0.833	6.271×10^{-9}	31) Fiihrer	0.782	1.991×10^{-7}
32) Reich	0.831	6.905×10^{-9}	32) Abetz	0.781	2.177×10^{-7}
33) somber	0.83	7.596×10^{-9}	33) Fascist	0.778	2.624×10^{-7}
34) planned	0.828	8.891×10^{-9}	34) Kertezsi	0.777	2.82×10^{-7}
35) Lanny	0.827	9.781×10^{-9}	35) Zaharoff	0.776	2.884×10^{-7}

Table 4.5: Highest ranked Spearman correlations for "Nixon" and "Spock" between 1960 and 1980. Required 60217 comparisons, $\bar{\alpha} = 8.303 \times 10^{-7}$. 752 significant correlations for "Nixon," 365 significant correlations for "Spock."

Nixon	ρ	p-value	**Spock**	ρ	p-value
1) Aldiss	0.968	8.333×10^{-13}	1) genres	0.964	2.409×10^{-12}
2) hardcover	0.968	8.333×10^{-13}	2) phaser	0.946	1.025×10^{-10}
3) nonfiction	0.964	2.409×10^{-12}	3) computer	0.942	2.003×10^{-10}
4) investigative	0.958	8.392×10^{-12}	4) communicator	0.938	3.641×10^{-10}
5) stomping	0.957	1.118×10^{-11}	5) generates	0.938	3.641×10^{-10}
6) Zelazny	0.956	1.587×10^{-11}	6) Afro	0.936	4.407×10^{-10}
7) cope	0.955	1.936×10^{-11}	7) programmed	0.926	1.779×10^{-9}
8) Moorcock	0.954	1.994×10^{-11}	8) paranoid	0.922	2.852×10^{-9}
9) fucker	0.952	3.25×10^{-11}	9) Aldiss	0.921	3.319×10^{-9}
10) tapes	0.952	3.25×10^{-11}	10) Uhura	0.92	3.527×10^{-9}
11) 1976	0.952	3.25×10^{-11}	11) Move	0.919	3.854×10^{-9}
12) manic	0.949	5.304×10^{-11}	12) helicopters	0.918	4.463×10^{-9}
13) prestigious	0.949	5.304×10^{-11}	13) film	0.917	5.157×10^{-9}
14) TV	0.949	5.304×10^{-11}	14) media	0.916	5.945×10^{-9}
15) vulnerability	0.949	5.304×10^{-11}	15) cinematic	0.914	6.837×10^{-9}
16) expertise	0.948	6.712×10^{-11}	16) operational	0.912	8.985×10^{-9}
17) paperback	0.948	6.712×10^{-11}	17) partially	0.91	1.027×10^{-8}
18) files	0.947	8.443×10^{-11}	18) worldwide	0.91	1.027×10^{-8}
19) 1940s	0.947	8.443×10^{-11}	19) printout	0.91	1.046×10^{-8}
20) bullshit	0.945	1.056×10^{-10}	20) counseling	0.909	1.171×10^{-8}
21) manipulate	0.945	1.056×10^{-10}	21) director	0.909	1.171×10^{-8}
22) options	0.945	1.056×10^{-10}	22) manipulated	0.909	1.171×10^{-8}
23) lifestyle	0.945	1.214×10^{-10}	23) manipulate	0.909	1.171×10^{-8}
24) briefing	0.944	1.314×10^{-10}	24) computers	0.908	1.333×10^{-8}
25) mainstream	0.944	1.314×10^{-10}	25) girlfriend	0.908	1.333×10^{-8}
26) format	0.943	1.626×10^{-10}	26) machine	0.908	1.333×10^{-8}
27) vulnerable	0.943	1.626×10^{-10}	27) adrenalin	0.906	1.515×10^{-8}
28) 1960s	0.943	1.626×10^{-10}	28) format	0.906	1.515×10^{-8}
29) upcoming	0.942	2.003×10^{-10}	29) Nixon	0.906	1.515×10^{-8}
⋮			⋮		
223) Spock	0.906	1.515×10^{-8}	223) Heritage	0.869	3.241×10^{-7}

Table 4.6: Highest ranked Spearman correlations for "lesbian" and "Picard" between 1980 and 2000. Required 47831 comparisons, $\bar{\alpha} = 1.045 \times 10^{-6}$. 1943 significant correlations for "lesbian," 805 significant correlations for "Picard."

lesbian	ρ	p-value	**Picard**	ρ	p-value
1) lesbians	0.988	5.504×10^{-17}	1) holodeck	0.956	1.336×10^{-11}
2) heterosexual	0.974	1.026×10^{-13}	2) Worf	0.953	2.664×10^{-11}
3) Feminism	0.97	3.812×10^{-13}	3) already	0.94	2.455×10^{-10}
4) politically	0.969	5.683×10^{-13}	4) relive	0.938	3.641×10^{-10}
5) family	0.968	8.333×10^{-13}	5) catalogs	0.931	9.098×10^{-10}
6) firsthand	0.956	1.478×10^{-11}	6) moot	0.929	1.28×10^{-9}
7) diversity	0.952	3.25×10^{-11}	7) pry	0.929	1.28×10^{-9}
8) Sexuality	0.951	4.165×10^{-11}	8) Losing	0.922	2.852×10^{-9}
9) culturally	0.949	5.304×10^{-11}	9) centerpiece	0.919	3.854×10^{-9}
10) Feminist	0.949	5.304×10^{-11}	10) hmm	0.919	3.854×10^{-9}
11) Homosexuality	0.949	5.304×10^{-11}	11) heading	0.918	4.463×10^{-9}
12) gay	0.948	6.712×10^{-11}	12) mimicking	0.918	4.463×10^{-9}
13) sexuality	0.948	6.712×10^{-11}	13) Riker	0.917	5.157×10^{-9}
14) civic	0.947	8.443×10^{-11}	14) So	0.917	5.157×10^{-9}
15) Kids	0.947	8.443×10^{-11}	15) trashed	0.917	5.157×10^{-9}
16) relationships	0.947	8.443×10^{-11}	16) locales	0.916	5.945×10^{-9}
17) pussy	0.945	1.056×10^{-10}	17) Crusher	0.913	7.846×10^{-9}
18) abuse	0.944	1.314×10^{-10}	18) caramel	0.912	8.985×10^{-9}
19) feminism	0.944	1.314×10^{-10}	19) Instead	0.912	8.985×10^{-9}
20) Politics	0.944	1.314×10^{-10}	20) bro	0.91	1.027×10^{-8}
21) status	0.944	1.314×10^{-10}	21) coordinates	0.91	1.027×10^{-8}
22) heterosexuality	0.943	1.626×10^{-10}	22) cuffed	0.91	1.027×10^{-8}
23) orientation	0.943	1.626×10^{-10}	23) shake	0.91	1.027×10^{-8}
24) tainted	0.943	1.626×10^{-10}	24) wishful	0.91	1.027×10^{-8}
25) laptop	0.943	1.675×10^{-10}	25) hovering	0.909	1.171×10^{-8}
26) appropriated	0.942	2.003×10^{-10}	26) Plus	0.909	1.171×10^{-8}
27) gender	0.942	2.003×10^{-10}	27) rotated	0.909	1.171×10^{-8}
28) homophobic	0.942	2.003×10^{-10}	28) vials	0.909	1.171×10^{-8}
29) baggy	0.94	2.455×10^{-10}	29) swatted	0.908	1.333×10^{-8}
30) ideologies	0.94	2.455×10^{-10}	30) tux	0.908	1.333×10^{-8}

I ("trenches" and "1918" are ranked 17th and 18th, respectively). When we focus on the period between 1905 and 1925 in Table 4.2, we see "war" more strongly correlated with World War I terms. For example, "Kaiser" is ranked 33rd in this list, and "Tommies," a slang term for British soldiers during World War I, is ranked 12th. Moreover, the highest correlation coefficient in this list is 0.926, which is much higher than the unfocused "war" correlations. When we test World War II in Table 4.3 by focusing on "war" between 1930 and 1950, the maximum correlation coefficient is the same as during World War I, and we see a lot of what we might expect to see in the top 20 correlations—e.g., "Nazis," "Hitler," "Fascist," and "Mussolini." We also see "planes" in 1st and "A.R.P." in 4th place, the latter being a reference to Air Raid Precautions, a British organization operating during World War II.

We list correlations for "Hitler" and "Lanny" in Table 4.4 for the period from 1930 to 1960. The highest correlation for "Hitler" is "Nazi" with a coefficient of 0.954. Ranked 2nd in this list is "Nazis." The maximum correlation coefficient for "Lanny" is much lower with 0.873 for "Detaze." (Marcel Detaze is Lanny Budd's step-father.) The second ranked correlation to "Lanny" in "Lannv," which is likely a optical character recognition error. "Budd" is 8th, which is encouraging. We see "Hitler" associated with many of the words in Table 4.3, as well as "Gestapo" and prominent Nazis like "Goebbels" and "Ribbentrop." Several words with the root "bomb" are also present. "Lanny" also appears and is ranked 35th in this list. For comparison, "Hitler" is ranked 12th among correlations with "Lanny." Although many of the words in the ranks for "Hitler" also appear in the ranks for "Lanny," which is not terribly surprising, since both words are most active during this period, the higher ranks for "Lanny" also include terms specific to the Lanny Budd series. Finally, we

note that "Hitler" has 105 signifant correlations, roughly twice as many significant correlations as "Lanny" (with 52).

When we compare "Nixon" and "Spock" in Table 4.5 for the period from 1960 to 1980, we see more science fiction than one might expect in the correlations to "Nixon." Specifically, "Aldiss" (rank 1), "Zelazny" (rank 6), and "Moorcock" (rank 8) are all science-fiction authors. The fact that these authors are mentioned in the data set at all is likely due to nonfiction works about fiction being labed as fiction [3] (SOM). Nonetheless, this is an interesting signal. We also see a few interesting words that appear more relevant to "Nixon." For example, "intestigative" and "tapes" are in the top 10. We also see "manic" at rank 10, "files" at rank 18, and "manipulate" at rank 21, along with some profanity sprinkled throughout. Unlike the previous example, the maximum correlation coefficients for both "Nixon" and "Spock" are about the same (0.968 and 0.964, respectively). However, similar to the last example, "Nixon" has more than twice as many significant correlations as "Spock." ("Nixon" has 752. "Spock" has 365.) The list of correlations for "Spock" unshockingly has references to "Star Trek." We see "phaser," "computer," and "communicator" between ranks 2 and 4. We also see "Uhura" at rank 10. Beyond these, the top ranks for "Spock" mostly resemble words correlated with "Nixon"—e.g., "manipulate" at rank 23. Similar to the last example, "Nixon" is ranked higher for "Spock" (29th) than "Spock" for "Nixon" (223rd).

Finally, we compare "lesbian" and "Picard" in Table 4.6 for the period from 1980 to 2000. The highest-ranked correlation for "lesbian" is "lesbians" with a correlation coefficient of 0.988. For "Picard," the highest-ranked correlation is "holodeck" with a slightly lower coefficient of 0.956. Again, we see more than twice as many significant

correlations for the exogenous feature, “lesbian” (1943), as we do for “Picard” (805). Unlike the previous example, the rankings for “lesbian” are unmistakably related to “gender” (27th) and “sexuality” (8th and 13th). We also see “heterosexuality” (2nd and 22nd) and “homosexuality” (2nd, 11th, and 12th), specifically, along with “homophobic” at rank 28. Furthermore, we see references to “Feminism” (3rd, 10th, and 19th). The meaning of “family” in rank 5 is ambiguous, especially considering “ideologies” in rank 30, however this only adds to the richness and relevance of this signal set. Unsurprisingly, “Picard” is correlated with several other characters from “Star Trek,” specifically “Worf” (rank 2), “Riker” (rank 13), and “Crusher” (rank 17). We also see “heading” (rank 11) and “coordinates” (rank 21). However, most of the signals associated with “Picard” are difficult to distinguish from normal language.

4.4 Concluding remarks

Side-by-side views of ranked correlations do not always yield obvious results—for “Nixon” and “Spock” (Table 4.5) in particular, the specific differences were fairly subtle. We do observe two qualititive differences between exogenous and endogenous features in the language. First, we tend to see more than double the number of correlations above the experiment-wide significance level for the exogenous word (e.g., “Hitler”) as for the endogenous word (e.g. “Lanny”). Second, we tend to see much higher maximum correlation coefficients for exogenous words than for endogenous words. Then again, we also saw that, in this respect, “Nixon” only beat “Spock” by a small margin.

It is worth noting that the ranks and correlation coefficients produced by this method are sensitive to the time period chosen. Expanding the boundaries of the time period in either direction risks introducing noise into the results. This may explain the small amount of ambiguity in the "Nixon" versus "Spock" rankings. It is also possible that "Nixon" simply is not exogenous in the same way that "Hitler" is exogenous.

The method we have used in this paper to explore the distinction between exogenous and endogenous features—between main plots and subplots—is not a one-size-fits-all technique. The time periods have to be chosen with some care, as do the words being compared over each given time period. However, our principled method of choosing words developed in [1, 2] also has the benefit of suggesting specific 2-decade periods over which to compare these choices, and our observations from the previous section suggest that a 21-year (and occasionally 31-year) correlation period can produce a fascinating variety of signals.

References

[1] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, "Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution," *arXiv preprint arXiv:1501.00960*, 2015.

[2] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, "Is language evolution grinding to a halt?: Exploring the life and death of words in english fiction," *arXiv preprint arXiv:1503.03512*, 2015.

[3] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

[4] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the google books ngram corpus," in *Proceedings of the*

ACL 2012 System Demonstrations, pp. 169–174, Association for Computational Linguistics, 2012.

[5] L. Jost, "Entropy and diversity," *Oikos*, vol. 113, pp. 363–375, 2006.

[6] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 1948.

Chapter 5

Conclusion

In the work presented in this dissertation, we have laid the foundation for future work involving the Google Books corpus. In Ch. 2, we demonstrated the risks involved in drawing overly hasty conclusions from even the most high-profile of data sets. In the cases of the unfiltered English data sets from both versions of the Google Books corpus and the English Fiction data set from the original version, we found clear signs of bias toward the inclusion of scientific texts. This bias must be considered when testing socio-linguistic hypotheses and publicizing results. Unfortunately, this bias is often overlooked in these contexts. Encouragingly, we were able to demonstrate that the 2012 version of English Fiction is not unduly affected by scientific literature, which makes this particular data set a useful benchmark for future research.

We also found consistently that the Google Books corpus weights contributions in favor of prolific authors (e.g. Upton Sinclair in the case of the "Lanny Budd" series) and to major franchises (e.g. "Star Trek" novels). This is to be expected in a data set that samples books and should be considered in contrast to data sets that more

directly reflect popular use of words and phrases. Nonetheless, this contrast is also often overlooked.

In Ch. 2, in addition to explaining the need for vigilance in employing high-profile linguistic corpora, we demonstrated an objective method for discovering words that are important to the dynamics in the Google Books corpus. We did this by making good use of information theory, specifically Jensen-Shannon divergence. In Ch. 3, we expanded this method to observe signals important to the flux of words across relative frequency thresholds. These observations, in turn, shed light on the differences between the signatures of endogenous effects, such as prolificity, on the corpus and exogenous effects, such as conflict. We explored this distinction in specific cases in Ch. 4. The results of this line of exploration invite future research into making good use of a lingustic corpus, such as Google Books, even if that corpus does not directly represent the popular use of language.

Finally, we note that our choices of words to be compared in Ch. 4 were guided by previously demonstrated principles. This improvement in the word selection process should help prevent the introduction of experimenter's bias in analyses involving specific groups of words. We have in fact supported the cause of principled analysis of linguistic data sets in general, and in doing so we have added to the landscape of socio-linguistic research.

Bibliography

[1] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

[2] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the google books ngram corpus," in *Proceedings of the ACL 2012 System Demonstrations*, pp. 169–174, Association for Computational Linguistics, 2012.

[3] J. M. Twenge, W. K. Campbell, and B. Gentile, "Increases in individualistic words and phrases in american books, 1960–2008," *PloS one*, vol. 7, no. 7, 2012.

[4] J. M. Twenge, W. K. Campbell, and B. Gentile, "Male and female pronoun use in us books reflects women's status, 1900–2008," *Sex roles*, vol. 67, no. 9-10, pp. 488–493, 2012.

[5] P. M. Greenfield, "The changing psychology of culture from 1800 through 2000," *Psychological science*, vol. 24, no. 9, pp. 1722–1731, 2013.

[6] A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley, "Statistical laws governing fluctuations in word use from word birth to word death," *Scientific reports*, vol. 2, 2012.

[7] M. Gerlach and E. G. Altmann, "Stochastic model for the vocabulary growth in natural languages," *Physical Review X*, vol. 3, no. 2, p. 021006, 2013.

[8] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc, "Languages cool as they expand: Allometric scaling and the decreasing need for new words," *Scientific reports*, vol. 2, 2012.

[9] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 1948.

[10] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951.

[11] J. Lin, "Divergence measures based on the shannon entropy," *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 145–151, 1991.

[12] R. A. Bentley, A. Acerbi, P. Ormerod, and V. Lampos, "Books average previous decade of economic misery," *PloS one*, vol. 9, no. 1, p. e83147, 2014.

[13] A. Koplenig, "The impact of lacking metadata and data truncation for the measurement of cultural and linguistic change using the google ngram datasets (draft - under review)," 2014. http://hdl.handle.net/10932/00-023C-DD02-76AF-FF01-9; accessed online January 5, 2014.

[14] Supporting Online Materials can be found at http://www.compstorylab.org/share/papers/pechenick2015a/.

[15] D. J. de Solla Price, *Little Science, Big Science*. New York: Columbia University Press, 1963.

[16] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, "Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution," *arXiv preprint arXiv:1501.00960*, 2015.

[17] Supporting Online Materials can be found at http://www.compstorylab.org/share/papers/pechenick2015b/.

[18] G. K. Zipf, "Human behavior and the principle of least effort.," 1949.

[19] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, "Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution," *arXiv preprint arXiv:1501.00960*, 2015.

[20] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, "Is language evolution grinding to a halt?: Exploring the life and death of words in english fiction," *arXiv preprint arXiv:1503.03512*, 2015.

[21] L. Jost, "Entropy and diversity," *Oikos*, vol. 113, pp. 363–375, 2006.