DHQ: Digital Humanities Quarterly
Preview
2023
Volume 17 Number 2

Computational Paremiology: Charting the temporal, ecological dynamics of proverb use in books, news articles, and tweets

Ethan Davis <edavis_at_lclark_dot_edu>, Computational Story Lab, Vermont Complex Systems Center, MassMutual Center of Excellence for Complex Systems and Data Science, Vermont Advanced Computing Core, Watzek Library, Lewis & Clark College
Christopher Danforth <chris_dot_danforth_at_uvm_dot_edu>, Computational Story Lab, Vermont Complex Systems Center, MassMutual Center of Excellence for Complex Systems and Data Science, Vermont Advanced Computing Core, Department of Mathematics & Statistics, University of Vermont
Wolfgang Mieder <wolfgang_dot_mieder_at_uvm_dot_edu>, Department of German & Russian, University of Vermont
Peter Sheridan Dodds <peter_dot_dodds_at_uvm_dot_edu>, Computational Story Lab, Vermont Complex Systems Center, MassMutual Center of Excellence for Complex Systems and Data Science, Vermont Advanced Computing Core, Department of Computer Science, University of Vermont, Santa Fe Institute

Abstract

Proverbs are an essential component of language and culture, and though much attention has been paid to their history and currency, there has been comparatively little quantitative work on changes in the frequency with which they are used over time. With wider availability of large corpora reflecting many diverse genres of documents, it is now possible to take a broad and dynamic view of the importance of the proverb. Here, we measure temporal changes in the relevance of proverbs within four corpora, differing in kind, scale, and time frame: Millions of books over centuries; thousands of books over centuries; millions of news articles over twenty years; and billions of tweets over a decade. While similar methodologies abound lately, they have not yet been performed using comprehensive phraseological lexica (here, *The Dictionary of American Proverbs*). We show that beyond simple partitioning of texts into words, searches for culturally significant phrases can yield distinct insights from the same corpora. Comparative analysis between four commonly used corpora show that each reveals its own relationship to the phenomena being studied. We also find that the frequency with which proverbs appear in texts follows a similar distribution to that of individual words.

Introduction

Our goal here is to advance “computational paremiology”: The data-driven study of proverbs, and in general to examine the utility of frequency-based studies of common phrases in large corpora. In particular, we hope to answer the following: Can a computational study of proverbs in large corpora offer unique insight in the study of those proverbs? And is this novel kind of approach appropriate to corpus linguistics and studies using corpora broadly? We first build a quantitative foundation by searching for and counting instances of an ecology of proverbs, and estimating their frequency of use over time in several large corpora from different domains. We then characterize basic temporal dynamics allowing us to address fundamental questions such as whether or not proverbs appear in texts according to a similar probability distribution to words [Zipf 2012] [Balasubrahmanyan 1996] [Williams 2015b] [Cancho 2001]. 1

In studies of phraseology, data on frequency of use is often conspicuously absent [Čermák 2014]. The recent proliferation of large machine-readable corpora has enabled new frequency-informed studies of words and *n*-grams (phrases of length *n*) that have expanded our knowledge of language use in a variety of settings, from the Google Books *n*-gram Corpus and the introduction of “culturomics” [Michel 2011] [Pechenick 2015a], to availability and analysis of Twitter data [Alshaabi 2020]. Routine formulae, or multi-word expressions that cannot be reduced to a literal reading of their semantic components, remain notoriously averse to reliable identification despite carrying high degrees of 2 symbolic and indexical meaning [Sag 2002]. It is, for instance, much easier to chart a probability distribution of single words or *n*-grams than phrases like proverbs, conventional metaphors, or idioms, which must be associated with a lexicon – a set of meaningful linguistic units.

Studies of words alone can generally assume that each word is lexically represented as such, but the study of phrases with a particular cultural use may require a lexicon in addition to the text being studied. Here, we use Mieder's *Dictionary of American Proverbs* as such a resource. Assuming words or "grams" as fundamental components of texts risks flattening the complex interrelations between common phrases that might place a text in a historical or literary context, by de-emphasizing or omitting altogether culturally significant phrases. We show a case (proverbs) in which established computational methods may be applied to a class of culturally meaningful phrases, with results that paint a valid and substantially unique picture of language use in the corpora studied. 3

Perhaps the most recognizable routine formulae are proverbs and their close cousin, idioms. Centuries of the study of proverbs – paremiology – have shown their importance in language and culture, and that they are immensely popular among the folk (the people for whom these phrases are culturally relevant) [Mieder 2012]. Proverbs are generally metaphorical in their use, and map a generic situation described by the proverb to an immediate context. In light of challenges in developing reliable instruments for measurement and quantification of figurative language, research would greatly benefit, as it has with words, from a better understanding of the frequency and dynamics of proverb use in texts. By applying new methodologies in measuring frequency and probability distributions, this study seeks to contribute to this endeavor. 4

Before going any further, we must detail a more precise definition of the proverb. Though there is still some debate, it is widely agreed that proverbs are popular sayings that offer general advice or wisdom. Naturally, not all such sayings are proverbs. Mieder's definition is perhaps the most useful for our present purposes: "Proverbs [are] concise traditional statements of apparent truths with currency among the folk. More elaborately stated, proverbs are short, generally known sentences of the folk that contain wisdom, truths, morals, and traditional views in a metaphorical, fixed, and memorizable form and that are handed down from generation to generation" [Mieder 2008]. 5

Proverbs maintain a particular relationship with their context of use that provides a fruitful domain for frequency and probability analysis. An important part of the proverb is the context in which it is used. The metaphorical property of a proverb need not only have to do with the proverb itself (as in the proverb/metaphor "war is hell," in which war is compared to hell within the proverb). In general, the use of a proverb is metaphorical in context, meaning that the proverb offers wisdom about a current situation via a metaphoric comparison to a proverbial one [Mieder 2008]. For instance, while the proverb "still waters run deep" might be used to caution someone against taking a seeming calm for granted, as it may belie unseen dangers. As with many other proverbs, it is hard to imagine anyone using the proverb "you can't put lipstick on a pig" in any literal or pragmatic context. Rather, these phrases offer wisdom embodied in the culture as opposed to that of the speaker. In this way proverbs may be used generically without proffering personal expertise. 6

Proverbs are necessarily ambiguous enough to offer wisdom in any number of situations. Michael Lieber argued that this ambiguity paradoxically gives proverbs the function of disambiguating situations in which they are used. In part due to their role as cultural rather than individual wisdom, they can be invoked impersonally as a way of clarifying a complex reality [Lieber 1994]. As such, part of Winick's definition of the proverb is that they "address recurrent social situations in a strategic way" [Mieder 2008]. 7

It is important to note the distinction between proverbs and idioms. An example of an idiom would be the phrase "red herring" denoting a mislead. The meanings of idioms, like proverbs, often cannot be ascertained from the meanings of their component words. But unlike proverbs, idioms are often not complete sentences, require context, and need not reference a paradigmatic situation. Proverbs on the other hand represent a complete situation and offer some sort of general wisdom. The boundary between the two is rather fuzzy and contains many idioms and proverbial expressions. For instance, the proverb "every cloud has its silver lining" is perhaps more well known by its idiomatic reduction "silver lining." In fact, people may use an idiom without any knowledge of its proverbial context. Our intent here is to focus on 8 expressions of full proverbs, and not their idiomatic uses. As previous work has shown, it is possible to investigate the manipulations and idiomizations of individual proverbs [Čermák 2014] [Moon 1998]. Our approach has limitations, and further research into flexible searches or other identification methods will be essential in future work.

Metaphor and idiom identification and comprehension are an open area of research in machine learning and NLP (Natural Language Processing) [Fazly 2009] [Shutova 2010]. In general, metaphors and metaphorical speech are difficult to identify, and do not occur in consistent, repeated phrasings. Whereas in the study of individual words, one is allowed the tacit assumption that most of these words are represented in the lexicon of the language, in the search for routine formulae, one must access the lexicon as an essential step in verifying a phrase's meaningfulness. Furthermore, the source and target domains of their metaphoric mapping are seldom explicit, as laid out by Lakoff and Johnson in their *Conceptual Metaphor Theory* [Lakoff 1985] [Andersson 2013]. Proverbs generally appear in the same recognizable format, and in the form of a full, self-contained sentence. Prospectively, understanding of the conceptual mapping involved in proverb use may provide a useful step towards general understanding of metaphors in the above fields [Özbal 2016a] [Özbal 2016b]. 9

Arguably, the proverb's flexibility of use has helped make them an essential part of language and communication, literature, discourse, and media [Mieder 2012]. Interest in the collection and study of proverbs dates back to at least the ancient Greeks and Sumerians. Erasmus famously collected proverbs. In English literature, the proverb has been an important device for many famous authors, among them Geoffrey Chaucer, William Shakespeare, Oscar Wilde, and Agatha Christie [Abrams 1994] [Obelkevich 1994]. 10

Quantitative Approaches

This is by no means the first quantitative study of proverb use. Permiakov called for demographic studies of proverb knowledge to gather an impression of which proverbs were being used by the folk, in the interest of establishing a paremiological minimum: A minimum lexicon of proverbs for understanding a language [Permiakov 1989]. Subsequent interest in proverb knowledge in psychology and folklore resulted in several studies conducted in the United States. Early studies by Albig and Bain in the 1930s found that American college students could recall on average between 25 and 27 distinct proverbs, many of which were common among participants [Albig 1931] [Bain 1939]. A more recent study by Haas observed proverb familiarity among college students in several regions of the US. They performed experiments in both proverb generation and proverb recognition. Notably, students could recognize more proverbs than they could recall on their own [Haas 2008]. 11

Apart from the lexicographic collection of proverbs from texts, several attempts have been made to quantify and characterize their use. Whiting, in his assiduous collection of proverbs from texts in "Modern Proverbs and Proverbial Sayings" [Whiting 2014], kept track of the frequency with which they were encountered. Norrick attempted a manual search for proverb frequency, though he was constrained to only using proverbs starting with the letter *f*, and used a relatively small text sample [Norrick 1985]. In the first serious computational analysis of proverb frequency, Lau searched for and counted instances of proverbs in newspapers in the Lexis/Nexis ALLNWS database [Lau 1996]. 12

David Cram theorized that proverbs, acting as self-contained lexical units, were employed much in the same way that words are, and that their use involved a "lexical loop" where the speaker accesses the lexicon in addition to the syntax when forming a text. As such, in the case of proverbs (and phrasal idioms), one ought to "analyze a syntactic string as a single lexical item" [Cram 1994]. 13

Moon's exhaustive early study of fixed expressions and idioms (denoted FEIs) in the Oxford Hector Pilot Corpus (OHPC) did just that [Moon 1998]. His study represents the first serious attempt to apply the new tools of computational linguistics to routine formulae. He searched the OHPC (a precursor to the British National Corpus or BNC) for instances of 6776 FEIs from the *Collins Cobuild English Language Dictionary*. It is worth noting that at the time, there were few machine-readable English phraseological lexica. Though proverbs consisted of only 3.5% of the searched phrases (240), 19% of the expressions found in the corpus were proverbial expressions, the second most common subtype behind "simple expressions" (70%). Of the proverbs found, 59% were deemed metaphorical. Moon notes that exploitation of FEIs are easy to miss, and uses the proverb "a bird in the hand is worth two in the bush" as an example. 14

Significantly, Moon noted that journalism was over-represented in the corpus, and that the results did not represent the distributions of these FEIs in English as a whole. This and other similar caveats inspired the present study to observe genre-specific corpora separately, and compare after analysis. 15

Čermák's essay collection *Proverbs: Their Lexical and Semantic Features* contains several essays that deal with the distribution of proverbs in the British National Corpus [Čermák 2014]. In Čermák's pioneering essays, he searches for occurrences of English proverbs in the BNC corpus (100 million words) [British National Corpus 2001]. In his study, even the most common proverbs seem to occur relatively infrequently. For example, "easier said than done" is the most common, appearing 62 times in the entire corpus. His study discusses the relevance of corpus occurrence to a paremiological minimum (he uses a limited proverb list from Wiktionary). Another study focuses on text introducers to various proverbs using collocation analysis. (Čermák notably created/spearheaded one of the first machine-readable phrasaeological lexica in the "Czech Idiom Dictionary" (1994).) 16

Čermák relates frequency dictionaries to discussions of a paremeological minimum. Should proverb frequency in large corpora be considered when judging that minimum? Of course, there are problems with this approach as well: proverbs rely heavily on oral tradition, and are prone to frequent corruptions and purposeful exploitations. As such there is no guarantee that a search of a given phrasing of a proverb will capture all, if any, of its occurrences in a text. There are ways around this on an individual basis, but it depends on the proverb: some employ parallel structures (like "good X make good Y"), or have popular idiomizations (like "silver lining"). Longer proverbs are more likely to appear in more than one form, as words and clauses can be swapped rearranged, or omitted without changing their overall meaning, which makes computational identification more difficult. Shorter proverbs will more reliably appear in a consistent form because there are fewer words to manipulate, and any manipulation is likely to change their meaning. 17

In a recent introductory paremiology textbook, Steyer (2015) outlined a process general corpus linguistic method for studying proverbs, similar to Moon and Čermák. Most recently, Haas (2022) used Google Trends to analyze the frequency of Google searches of proverbs included in discourse around the COVID-19 pandemic. Here, we expand on the above literature, including much larger corpora and proverb data sets. 18

Should the ambition be to find these distributions in English as a whole? We contend that there is no such universal corpus for any language. Clearly use of these phrases is context-dependent, it seems unlikely inter-contextual searches will yield greater insight than single-genre searches. Instead, frequency dynamics and distributions in separate corpora from differing contexts may be more informative. 19

From Data on Language to Culture

Our present study of proverbs from a corpus linguistic point of view examines proverb frequency using several methods that are well established in the study of words, though rarely used in conjunction: namely the dynamics of frequency over time, and the relationship between frequency and relative popularity (rank). We illustrate that moving beyond words to the study of significant phrases provides worthwhile insight that cannot be captured by word-based methods, and yet reproduces some of the expected behavior of words. 20

One of the foundational achievements in the study of complex systems was Zipf's identification of scaling laws in language and other social phenomena [Zipf 2012]. Studies of scaling can help us understand the relationship between common and uncommon observations, and are particularly suited to studies of language, which relies both on the utility of some elements, and the specificity of others. A rank distribution describes the relationship between the frequency of a word's appearance, and its resulting rank among all words in the text. Zipf's law shows that there is an inverse relationship between the frequency and rank of words in a text, and that the most common 20% of the unique words in a text account for 80% of overall word frequency. It was first observed by Zipf that the rank distribution of words in a text follows a *power law* $F(r) = cr^{-\alpha}$, where r is a word's rank, $F(r)$ is its frequency, with $\alpha \simeq 1$. As early as 1996, natural language (in the context of computational linguistics) was cited explicitly as an example of the recently coined "complex adaptive systems" [Balasubrahmanyan 1996]. 21

While primary interest here is paid to its appearance and seeming ubiquity in language, the same class of distributions have been observed in phenomena across a wide range of fields including physics, biology, psychology, sociology, urban studies, and engineering [Clauset 2009] [Martínez-Mekler 2009]. 22

One shortcoming noted in many evaluations of Zipf's law in text is that power law scaling breaks down toward the tails of these empirical distributions, meaning that the lowest ranked words do not seem to follow Zipf’s law. Recent work by Williams et al. [Williams 2015b] showed that power law scaling holds over more orders of magnitude when randomly partitioned phrases are used rather than individual words. That study also suggested a refocusing of corpus linguistic attention from words to phrases as essential elements of language. Further work by Williams et al. (2015a) suggested that changes in scaling in Zipf distributions of large corpora can be attributed to text mining. Few, if any, attempts have been made to apply Zipf's law to phraseological lexica. 23

With large amounts of newly digitized text, corpus linguistics and lexicology/lexicography have seen renewed wider interest, and new results. Can these methods be used to tell new stories that are of interest to those working in the humanities? And in particular, how can that work embed itself into the existing wealth of knowledge accrued by those disciplines. In this case, how can computational work on proverbs situate itself in the existing knowledge-base of paremiology? 24

In their seminal 2011 paper, Michel et al. discussed the newly created Google Books corpus, and coined the term “culturomics” to describe the nascent discipline concerned with observable trends in the use of *n*-grams over time [Michel 2011]. They present several case studies, among them trends in the use of “influenza” with historical outbreaks, and the use of geographical and antagonistic terms alongside the history of the American Civil War. These case studies make use of time series data and relative frequency to tell complex stories of interest from simple queries. 25

Pechenick et al. note that there are serious issues with assertions that Google Books offers a reliable representation of culture. For one, books are not indexed by popularity, and each book appears only once. As a result, the linguistic contributions of the most popular books are weighted equally with the least popular [Pechenick 2015a]. Secondly, the increase in volume of scientific publications in the last century causes the last century of English as a whole to be relatively skewed towards that genre. For instance, enormously influential books like *To Kill a Mockingbird*, *I Know Why the Caged Bird Sings*, *Mockingjay*, or *Harry Potter and the Order of the Phoenix* are only represented once, and share the same weight as even the most obscure books. In the last century, the rise in volume of scientific and academic publication drastically increased the relative influence of this type of writing. Here, we examine only the English Fiction subset of the corpus, which is a less problematic subset [Pechenick 2015b]. 26

Other work by Reagan et al. utilized the timelines *within* texts to evaluate the “emotional arc” of a text, given word valence (sentiment) data. Emotional arcs were created by plotting the changing sentiment of words in a text as it progresses: from beginning to end, how positive or negative is the overall languge in a given section? Inspired by Kurt Vonnegut's rejected Master's thesis (in anthropology) on the shapes of stories, they found that the emotional arcs of most stories in the Gutenberg corpus could be reduced to a handful of paradigmatic shapes [Reagan 2016]. 27

Work by Underwood et al. used historical use of gendered names and words to reveal trends in gender representation in literature using data from the HathiTrust digital library [Underwood 2018]. 28

StoryWrangler, a tool recently developed by Alshaabi et al. allows users to explore the temporal dynamics of *n*-grams found on Twitter [Alshaabi 2020]. Using a data set reflecting a random 10% of Twitter since 2008 (presently over 150 billion tweets), Storywrangler tracks the prevalence of *n*-grams on a daily scale. *n*-grams are portrayed via rank by popularity, and convey the rise/dynamics of President Trump (further depicted in the PoTUSometer) [Dodds 2021], or the meteoric rise, and continued influence of Justin Bieber (of surprising relevance to this work). Unlike the Google Books *n*-gram Corpus, StoryWrangler is notable in its ability to track phrases in both original tweets and retweets, conveying aspects of popularity through amplification. 29

Beyond simple words and phrases, data have been used to track the progression of ideas. For instance, Leskovec et al.'s paper on “meme-tracking” tracked the progression and mutation of popular sayings as they proliferated through 30 news reporting and blogging [Leskovec 2009].

Recently, “Computational Folkloristics” has gained recognition as an area of study, with a 2016 issue of the *Journal of American Folklore* being devoted to the subject [Tangherlini 2016]. Using classification, networks, geographical data, temporal data, and digitized text, folklorists and other interested academics have explored new possibilities in understanding texts and cultural history. The *Danish Folklore Nexus* developed by Abello et al. provides tools for large-scale analysis of Danish folk tales and stories, aiding in classification of stories, or mapping their similarity to others through networks. Tools like this can augment traditional methods of studying folklore, using data-driven methodology to guide future avenues of folklore research [Abello 2012]. This represents a paradigmatic example of a computational tool participating in the continued discourse around folklore, without being an end in and of itself. 31

Data and Methods

In an effort to quantify the ecology of proverbial language, a list of over 14,000 proverbs was obtained from Mieder's *Dictionary of American Proverbs* [Mieder 1992]. Proverbs were stored in an SQL database for ease of access, and matched for frequency with four distinct corpora: 32

- The Gutenberg Corpus (English)
- The *New York Times* (1988-2007)
- The Google Books *n*-gram Corpus (1800-2000)
- Twitter (2008-2020)

For Google Books, the proverb “shit happens” was added to the set of proverbs to illustrate the emergence of a modern proverb. 33

Individual corpora were collected as follows. 34

A. Gutenberg

The Gutenberg corpus comprises over 60,000 collected published documents spanning several centuries. The present study restricts its use to the subset of documents in English. As the metadata for the Gutenberg corpus does not consistently encode the date of original publication, temporal data was collected using author birth dates (gathered from the gutenbergr library for R) [Robinson 202]. These were used in place of publication dates, as the publication dates in the corpus seldom represent the original publication, instead they represent the digitized edition. For temporal analysis, documents without authors and their birth dates were omitted. 35

The Gutenberg corpus comes with several caveats. Firstly, works were curated by perceived importance. Works also disproportionately represent the 18th and 19th centuries, and for this reason much of our work with Gutenberg focuses on this period. Several authors have much of their extensive oeuvre represented in the corpus (e.g., Anthony Trollope, Mark Twain), which could compromise a more objective view of English writing tendencies of the period. 36

B. The *New York Times*

Data from the *New York Times* were gathered from the *New York Times* Annotated Corpus of 1.8 million articles from 1987-2007 [Sandhaus 2008]. The data are organized in NTIF (News Industry Text Format) formatted XML-readable documents. The corpus includes obituaries and other short pieces in addition to more traditional news articles. 37

C. Google Books

The 2020 English Fiction Google *n*-grams corpus consists of every *n*-gram that appears at least 40 times in its set of millions of digitized books. For each *n*-gram the corpus provides on each year it appears in the data set, the frequency with which it appeared that year, and the number of documents it appeared in that year [Michel 2011]. Unlike Gutenberg and the *New York Times*, Google Books does not contain the raw text of the concerned documents, rather it displays the counts for popular *n*-grams in the corpus, organized by *n*-gram length. This creates an obvious limitation, where only 38 phrases of the same length can be studied together.

Twitter

Data from Twitter was accessed through the Vermont Complex Systems Center's StoryWrangler API [Alshaabi 2020]. StoryWrangler receives a randomly selected 1/10th of each day's tweets from Twitter's Decahose API (including retweets), and organizes *n*-grams by rank and frequency. Data for 2-gram and 3-gram proverbs were obtained though the tool, and were aggregated so the collection was case insensitive. Similar to Google Books, this corpus is organized by *n*-gram counts rather than full texts. 39

D. Data Processing and Visualization

The data from all four corpora were processed using Python, and the libraries pandas and matplotlib were used for organization and visualization respectively [pandas 2020] [Hunter 2007]. 40

In our processing of Gutenberg and the *New York Times*, punctuation in both proverbs and texts was removed. Twitter data were punctuation insensitive. Regular expressions were used to capture variations in punctuation when processing the Google Books *n*-gram Corpus. 41

Estimating the frequency of a proverb’s use is essential to this work. Simple counts don’t lend themselves to comparing results between corpora of different sizes, and do not capture the frequency of the proverb in relation to the size of a single corpus. Relative frequency can be calculated simply by dividing the raw frequency by a quantity describing the size of a corpus (number of articles, number of *n*-grams, number of books, etc.) Expressed mathematically, it is calculated as: $f_{rel} = f_t / n_t$ which is the frequency f for time period t divided by the number of documents n found during time period t. 42

Zipf distributions were plotted using ranks of proverbs in a corpus, with rank 1 being the most frequent against their frequency. Zipf distribution plots are shown on a log-log scale as is standard, which displays less intuitive power law functions as more intuitive linear functions. For results and analysis, see Appendix A. 43

A network approach will be useful in exploring how different books are connected by the proverbs they share. In this case, a connection is drawn between two books if they share at least one proverb. The resulting network emerges after this step is performed for each possible pair of books. In network analysis, an important metric is *centrality*, which in this case describes how well-connected a book is in the larger network. Specifically, we calculate *betweenness centrality*, which assesses how often a given book appears in the shortest path between any other two books in the network. A book with high betweenness centrality in this network appears in the path between many pairs of books in the network. For results and analysis, see Appendix B. 44

Betweenness centrality in these networks is calculated as $b(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$,where σ_{st}denotes the number of shortest paths between books *s* and *t*, and $\sigma_{st}(v)$denotes the number of those paths that also pass through book *v*. 45

Most processing was performed using the Vermont Advanced Computing Core (VACC) located at the University of Vermont. 46

Results

Gutenberg

While the most popular entry in the Gutenberg corpus and the Google Books *n*-gram corpus was the phrase “hold your tongue,” this phrase is classified as a proverbial expression rather than a proverb (its use requires outside context). For clarity of focus the phrase has been excluded from figures in this section. “Sink or swim,” another proverbial expression, has been left in. In light of the limitations of the Gutenberg corpus detailed in Methods, it is difficult to make claims about the trends of proverb use over time (Figure 1). It is clear from the data shown in Figure 1 that proverbs appear in a remarkable portion of the documents in the corpus. “The sooner the better” for example, appears in nearly one in every 47 ten documents in the early 1800s.

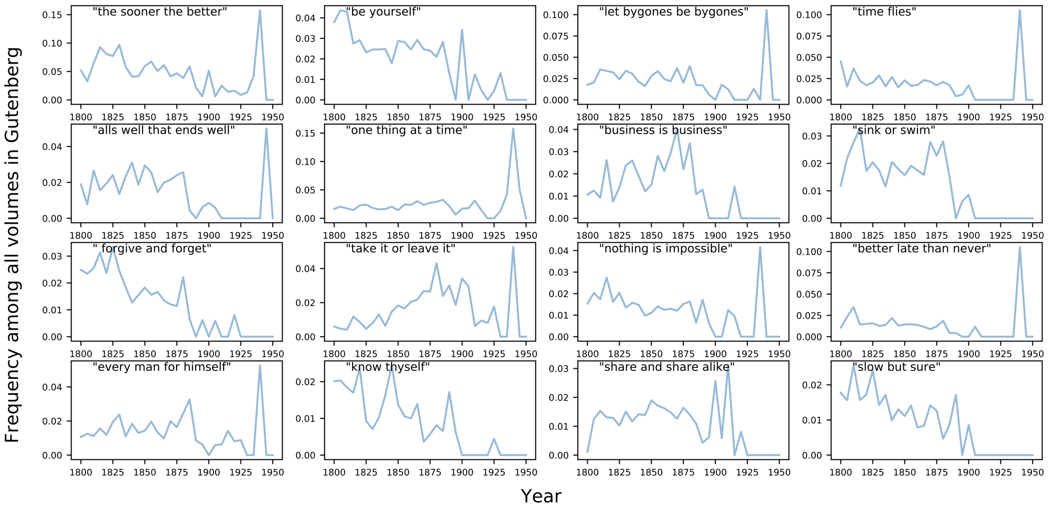


Figure 1. Time series for the 16 most popular proverbs in the Gutenberg corpus, ranked by overall count. These most common proverbs occur in a large portion of documents in the corpus for most of the period studied. For instance, “the sooner the better” regularly appeared in at least 5% of documents from the 19th century. Plots are ordered in the grid by rank first left to right, then top to bottom. Note that the vertical axis ranges vary across plots to highlight individual variation in time.

The *New York Times*

Figure 2 shows time series plots for the 16 most common proverbs in The *New York Times* Annotated Corpus. Shown are frequency binned by month and year, and normalized by article count. All articles are included in the count including smaller articles like obituaries (the average article count is 248 per issue). It is by no means a surprise that proverbs appear frequently in journalism; in fact Lau's study found as much [Lau 1996]. Not present in that work, is a temporal dimension (not to mention a different time period). It is clear in Figure 2 that the proverbs represented are used on a monthly or semi-monthly basis, and are rarely if ever absent in a year's publications. In these representations of proverb use, it is easier to identify use patterns and perhaps to extract narratives from their dynamics. The easiest, if somewhat trivial case is “to delay may mean to forget” owes its yearly rhythm to its role as the NYT's charity tagline. Its frequency of use increased markedly over the period studied, though stayed confined to the winter holiday months.

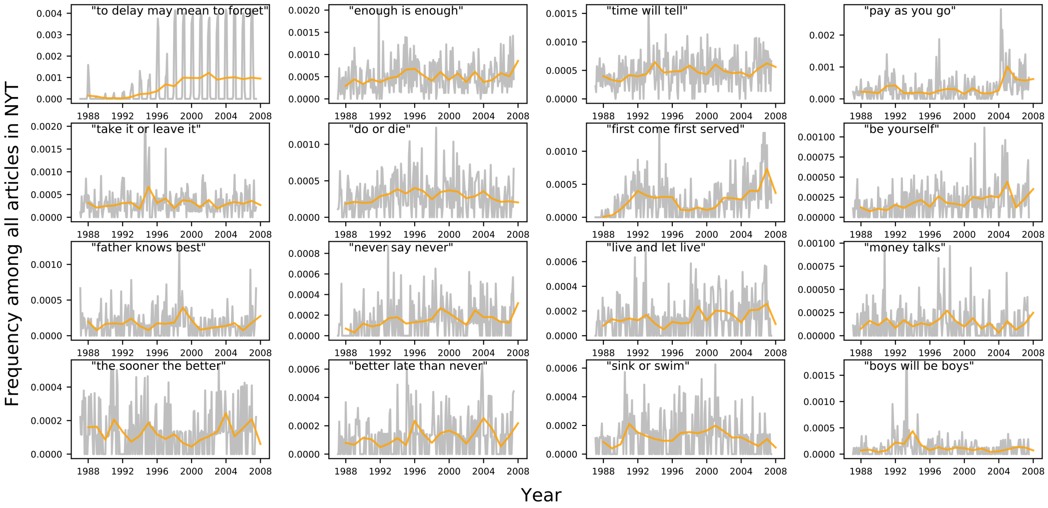


Figure 2. Time series plots for the 16 most popular proverbs in the *New York Times* from 1997-2007 (ranked by overall count). The gray represent the data binned by month, and the orange represent the data binned by year. The proverb “to delay may mean to forget” owes its yearly rhythm to its role as the NYT's charity tagline. The frequencies are normalized by article count (obits, and non-body included). Plots are ordered in the grid by rank first left to right, then top to bottom.

With the exception of “to delay may mean to forget,” and consistent with accepted definitions of the proverb, the consistency with which proverbs are used in the *New York Times* suggests they are employed widely for their utility in mapping general wisdom to a specific context. 49

Nonetheless, prominent spikes in frequency can be associated with historical events. For instance, the brief several-fold increase in the use of “boys will be boys” around November of 1992 is likely attributed to a contentious and widely publicized sexual assault case at the time, which prompted additional discussion of rape culture. Before the trial, the president of the New Jersey chapter of the National Organization for Women was reported as saying, “We're going to stop this 'boys will be boys' attitude from continuing in this country.” Meanwhile in the trial, the lawyer for the defense excused the rapists’ actions, telling the jury, “boys will be boys” [Glaberson 1992] [Hanley 1992]. 50

The maximum in use of “pay as you go” in 2004 seems to correspond with discussion around President Bush and the Republican party’s budget plans, referencing the “pay as you go” budget policy from the 1990's by which tax cuts and spending increases must not increase the defecit. Its increase in use in 1996 seems to owe to discussion of the Environmental Bond Act being proposed in New York at the time, which proponents argued would cause fewer delays than “pay as you go” funding for environmental clean-up [Vote Yes on the Bond Act 1996] [Henry 1996]. 51

Google Books

In Figure 3 are time series plots for the 12 most common 2-gram proverbs in the Google *n*-grams corpus. Here the gray represents yearly frequency (counted once per volume), and the orange represents the five-year rolling average, normalized by the number of volumes in a given year. One can see clearly from the figure the emergence of several more recent proverbs: “safety first,” “money talks,” and “shit happens.” 52

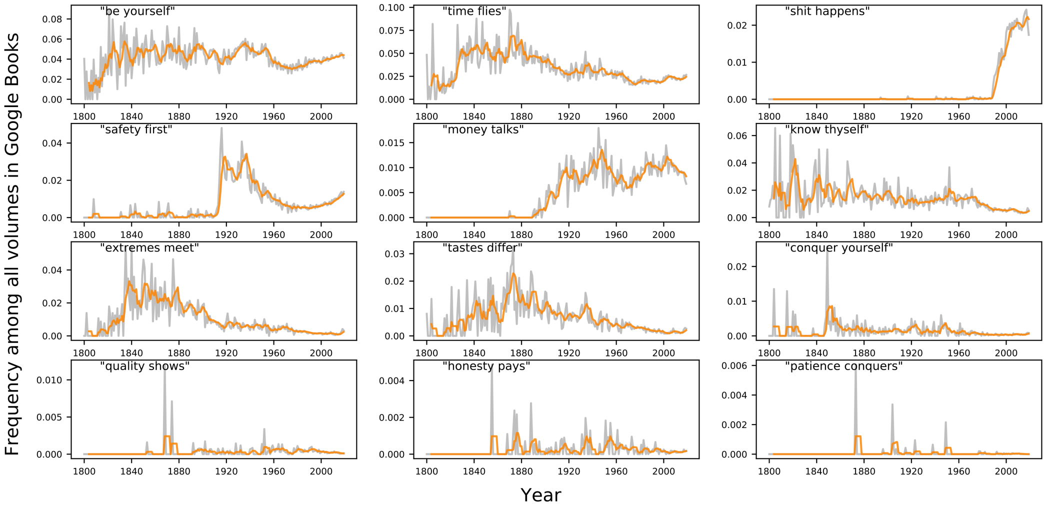


Figure 3. Time series plots for the 12 most popular 2-gram proverbs in the Google Books n-gram Corpus (ranked by overall count). The gray represent the yearly frequency, while the orange represent the five-year rolling average. The dramatic increase in use of the proverbs “shit happens” and “safety first” correspond with previous scholarship on their emergence. Plots are ordered in the grid by rank first left to right, then top to bottom.

53 “Safety first” exhibits a precipitous rise in usage in the early 20th century. Specifically, in 1912, the National Safety Council (NSC) in the US adopted the phrase as its slogan to promote standards of worker safety, though the Safety First Movement was initiated by US Steel in 1906. Its origin has been traced back to at least 1818 [Mieder 2019]. The data shown in Figure 3 support the history of its popularization [Swuste 2010] [Mieder 1992].

54 Previous scholarship on the proverb “shit happens” (which does not appear in *The Dictionary of American Proverbs*) traced its origin to the year 1944, and its rise in popularity corresponds to its humorous use as a bumper sticker, and cultural controversy (and legal battles) associated with it [Mieder 2004] [Georgia 1991]. It’s increasing currency is illustrated in its appearance in Tom Clancy’s popular novel, *Clear and Present Danger*: “Look, in field operations anything can go wrong … We are not immune. Shit happens, as they say” [Clancy 1990]. It also famously appeared in the movie *Forrest Gump* [Zemeckis 1994].

55 Figure 4 shows time series plots for the 16 most popular 3-gram proverbs in the Google Books *n*-gram corpus. Though the proverb “never say never” originated in 1887 [Mieder 1992], it is evident that it gained far wider popularity in the late 1900s. Though the proverb “enough is enough” dates at least to 1546 [Mieder 1992], its popularity seems to vastly increase throughout the 20th century. The proverb “divide and conquer” seems to have briefly gained popularity around the World War II era.

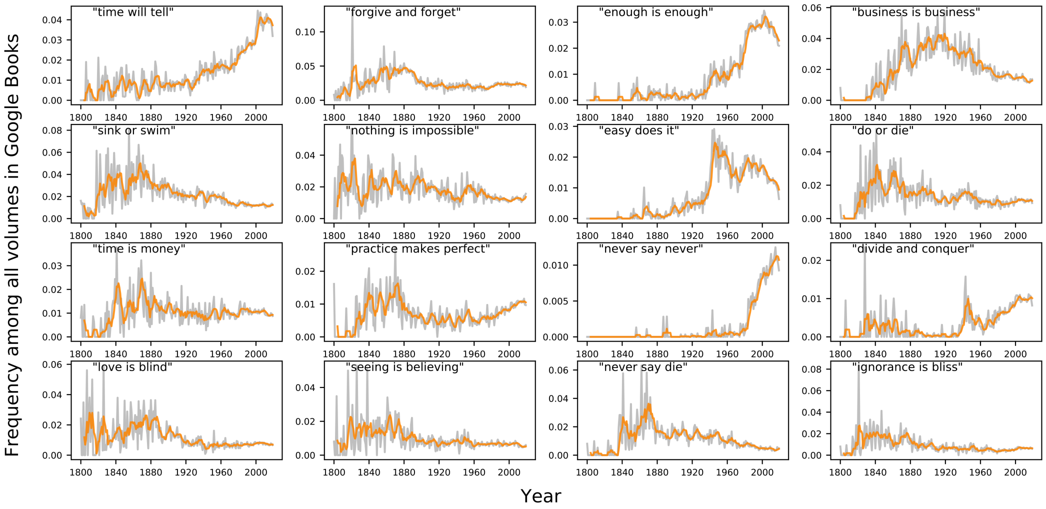


Figure 4. Time series plots for the 16 most popular 3-gram proverbs in the Google Books Ngram Corpus (ranked by overall count). The gray represents the yearly frequency, while the orange represents the 5 year rolling average. The rise in popularity of the proverb "never say never" is shown. A period of increased usage of the proverb "divide and conquer" corresponds with the World War II era. Plots are ordered in the grid by rank first left to right, then top to bottom.

Twitter

56 On Twitter, the four most common 2-gram proverbs, on average, don't seem to exhibit much variability in their usage (Figure 5). The proverbs "be yourself" and "time flies" seem to remain above 10^{-6}, or 1 in every million 2-grams on Twitter during the period studied. An increase in usage of "safety first" in early 2020 may be related to the onset of the coronavirus pandemic during the same period.

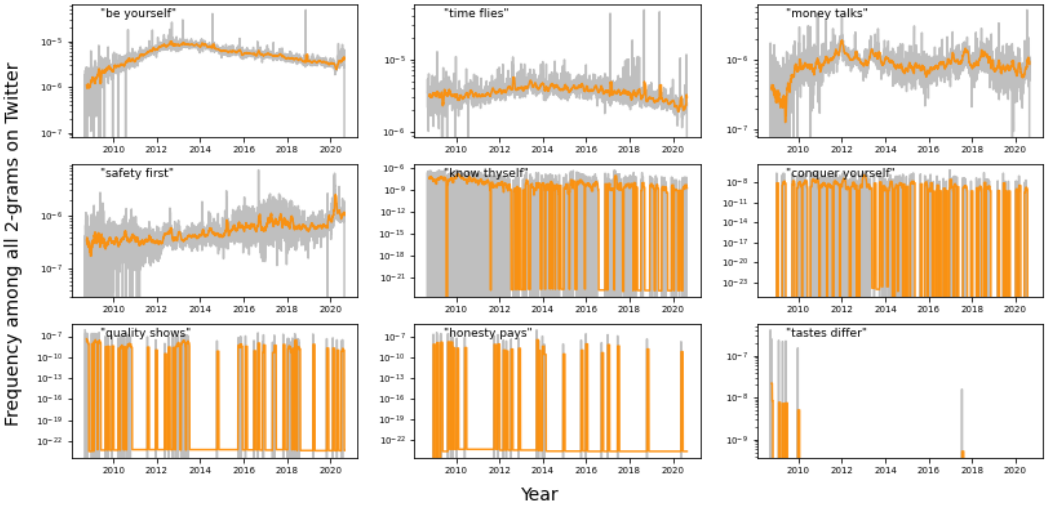


Figure 5. Time series plots for the nine most popular 2-gram proverbs on Twitter (ranked by overall count). The gray represents the daily frequency, while the orange represents the 30 day rolling average. The proverbs "be yourself" and "time flies" maintain popularity over the period studied. Notably, the "safety first" shows an increase in popularity in early 2020, possibly relating to the coronavirus pandemic. Plots are ordered in the grid by rank first left to right, then top to bottom.

57 Exhibited on Twitter (Figure 6), the convenience of proverbs as succinct narratives has made them useful in several titular media events in the past decade. Of note, Figure 6 shows marked shifts in frequency of "never say never," and "love is blind." "Never say never" owes its initial attention in 2010 to Justin Bieber's single of the same title (*Justin Bieber: Never Say Never*), repeated as his slogan and title of a biographical documentary. This was not the first film to

utilize the proverb in its title; Sean Connery's final performance as James Bond was titled *Never Say Never Again* (1983).

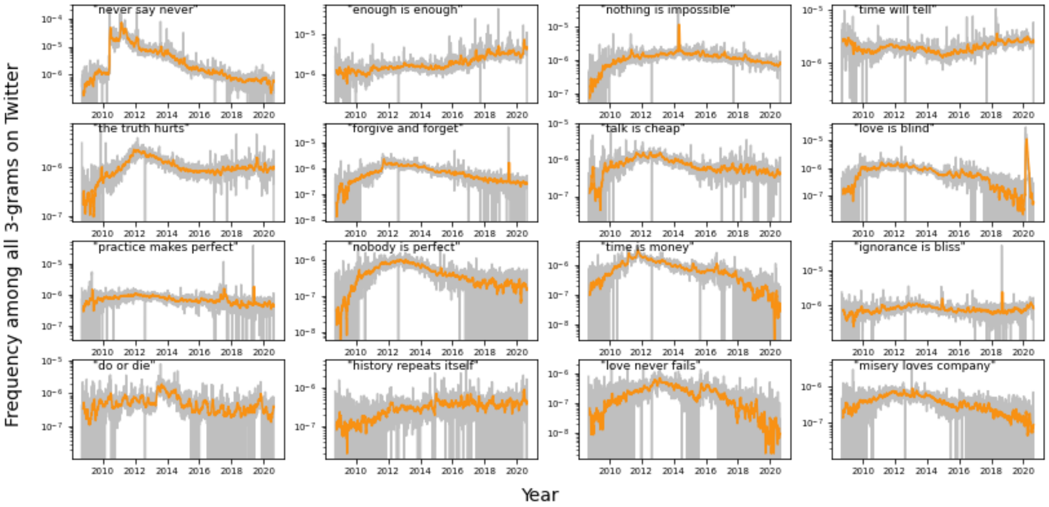


Figure 6. Time series plots for 3-gram proverbs on Twitter (ranked by overall count). The gray represents the daily frequency, while the orange represents the 30 day rolling average. The proverb “never say never” owes its meteoric rise in popularity in 2010 to popular musician Justin Bieber's single and biographical documentary of the same name. “never say never” remains the most popular proverb on Twitter until 2016, when it is supplanted by “enough is enough” which has steadily gained popularity in the last decade, owed in part to its constant use by Senator Bernie Sanders, and punctuated by reactions to tragedies related to gun and police violence. Plots are ordered in the grid by rank first left to right, then top to bottom.

Figure 7 shows the dynamics of “never say never” on Twitter in more detail. We observe first its meteoric rise in popularity at the time of *Never Say Never*'s (song) release as the lead single off the soundtrack for a modern remake of the *Karate Kid* movie (roughly two magnitudes in a single day). At the time of the single's official release on June 8th, 2010, “never say never” was the 63rd most used 3-gram on Twitter. When *Justin Bieber: Never Say Never* was released on January 31, 2011, “never say never” was the 34th most common 3-gram on Twitter; for comparison, “I love you” was 22nd at the time.

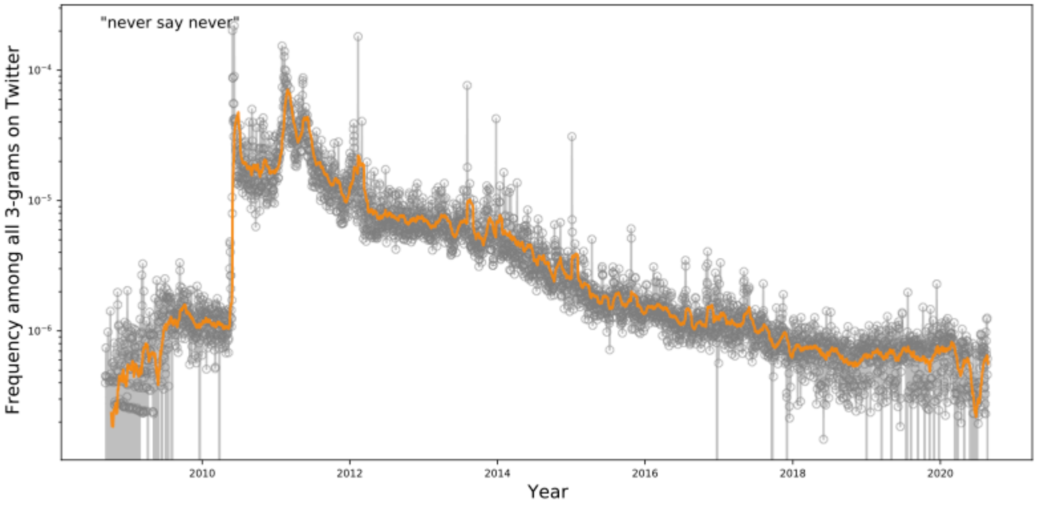


Figure 7. Daily relative frequency of the 3-gram “never say never” on Twitter. While “never say never” was already popular on Twitter as of 2008, its popularity was amplified in 2010 by the release of Justin Bieber's single entitled “Never say never,” and his subsequent biographical documentary of the same name. Remarkably, it remained the most popular proverb on Twitter for almost six years, punctuated by anniversaries and reruns of the movie, until it was surpassed by “enough is enough” in 2016.

59 Remarkably, the popularity of “never say never” on Twitter decayed so slowly that it did not reach its pre-Bieber frequency until 2016. The continued presence of the proverb in Twitter discourse suggests that in the wake of its initial rise, it was more frequently adopted to general non-Bieber usage. (A similarly popular 3-gram, non-proverbial song of that year, “Rock That Body” appeared and disappeared from the Twitter discourse in the span of a few months). While the enormity and fervor of Bieber's fanbase at the time (a period called “Bieber fever” [Tweedle 2012]) certainly contributed to its popularity, its continued use over a five-year period is compelling evidence that the proverb became a more integral part of the Twitter lexicon for a time.

60 In 2020, “Love is Blind” became the title of a literally minded reality dating show in which participants were quarantined in private rooms, only communicating via audio interfaces. In this instance, the proverb was not only an apt description of the show's narrative, but a template for its formation. Additionally, it came to represent a narrative solution to the isolation imposed by the concurrent pandemic. The increase in the phrase's popularity seems only to have lasted for the month of the show's release, after which it seems to settle at its former rate of use. The proverb itself is ancient, and translations exist nearly every European language.

61 While with “never say never” (the most popular proverb on Twitter), we see a sudden rise and slow decay, we see a different pattern in the second most popular proverb, “enough is enough.”

62 From 2016 to the present, we see a steady increase in the frequency of “enough is enough” on Twitter (Figure 8). Recent work by Mieder attributes its renewed popularity in part to its constant use by Bernie Sanders [Mieder 2019]. Unlike “never say never” there does not seem to be a single event that precipitates this trend. An investigation into the several local maxima (brief spikes in occurrence) suggests a possible narrative correspondence. Many of these local maxima correspond to events related to either police violence or mass shootings.

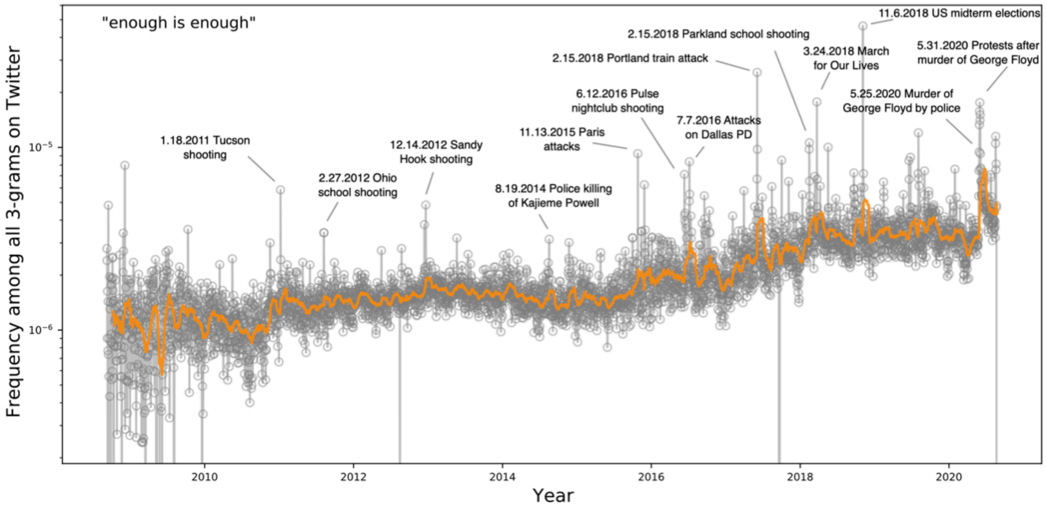


Figure 8. Daily relative frequency of the 3-gram “enough is enough” on Twitter. The popularity of “enough is enough” on Twitter grew steadily over the last decade, and it has been the most popular proverb on Twitter since 2016, perhaps originating from its consistent use by Senator Bernie Sanders [Mieder 2019]. It has since become associated with growing protests against police brutality and gun violence. Annotations reflect widely reported violent events and protests (with the exception of the 2018 US midterm elections). The stark simplicity of this sixteenth century proverb evokes a narrative of repetition past the point of tolerance [Mieder 1992]. In this instance, beginning as a condemnation of the continued reaffirmation of the status quo in US politics by Senator Sanders, it is now popular as collective outcry against political inaction in the wake of regular mass shootings in the US, and a lack of accountability in the killing of black Americans by police. The changing significance and popularity of the proverb in the past decade displays the aptitude of proverbial speech to be successfully employed in varying contexts, and its potential to illustrate narrative commonalities between phenomena.

63 Famously, survivors of the Parkland shooting in 2018 appeared on the cover of *Time* magazine with a simple title: “Enough.” [Alter 2018]. Coverage of the March for Our Lives against gun violence in the *New York Times* included the title: *March for Our Lives Highlights: Students Protesting Guns Say “Enough Is Enough”* [March for Our Lives 2018]. When protesters marched in DC in the wake of the murder of George Floyd, Politico's coverage was titled: *“Enough is*

enough": Thousands descend on D.C. for largest George Floyd protest yet [Semones 202]. Inasmuch as proverbs can create metaphorical mappings from a paradigmatic situation (or narrative) onto a present one, "enough is enough" represents a compelling narrative of continued injustice, and a critical point of retaliation. The data from Twitter display a narrative of repeated tragedy in spite of public outcry. The proverb was most popular during the 2018 US midterm elections, during which gun control was a major issue.

Concluding remarks

In this study, four corpora reveal four markedly different patterns of proverb use, which taken each within the limits of their methods of collection, can offer piecemeal insight into the relationship between phrases and the written record. Observing all four side by side, it should become clear that generalizing linguistic tendencies from individual corpora may not be a good idea. Even among corpora with significant temporal overlap the results differ due to both the scope of each corpus, and the way in which data are collected from texts. The present approach of studying data from distinct domains allows for both a more limited and more useful interpretation of the results: We can only claim that results are representative of proverb use on Twitter for instance, rather than proverb use in English as a whole — an impossible achievement. 64

So, rather than using the results to speak about *all proverbs* or all of language, we may use it as lens by which to identify areas (within a corpus) that warrant a closer look. In studies of common words, the concept the word represents is often dependent on its context, and evaluating context for each instance in a large corpus is generally not feasible. Proverbs carry with them a paradigmatic context 65

Unlike studies of common words, whose interpretation is often dependent on their context, we show that proverbs allow us to perform a different kind of analysis that focuses on the historical use of a stable paradigmatic narrative. 66

While much work up to this point uses computational methods to take a broader view of text than traditional methods of humanities scholarship, that broader view often necessarily obscures the context in which linguistic elements are used. A single word or short sequence of words may carry any number of meanings dependent on the words or concepts around it. Proverbs carry their own paradigmatic context which makes their meaning and use fairly consistent. While proverbs may be employed in many different practical situations, the paradigmatic situation they represent remains the same. In searching for proverbs, we are able to see how a paradigmatic context is applied to different practical contexts. The results of this search show not only the frequency of some sequence of words, but also the frequency with which a particular cultural concept is employed. 67

N-gram based methods of analyzing common phrases in texts suffer from the inclusion of many sequences of words which do not have any self-contained meaning. Our work, in which phrases must be represented in the lexicon, ensures that each phrase studied is meaningful. In an analysis of all 3-grams in a text, a very common phrase without a fixed meaning, like "and this is" could obscure a far less common meaningful phrase like "love is blind," even though it is common among phrases with fixed meanings. Our study of a specific cultural-linguistic phenomenon highlights the utility of using an extensive lexicon to isolate the phenomenon being studied. 68

We demonstrate that lexically significant phrases (here proverbs), which are relatively stable and self-contained, can discern trends that words or *n*-grams would likely miss. In each corpus studied, evidence of proverbs' changing use over time is shown to validate previous scholarship, reflect cultural events, and offer a quantitative, longitudinal perspective unable to be achieved by traditional methods in the study of proverbs. 69

The study of common and flexible phrases seems to lend itself to this mode of inquiry. Through novel or context-specific words and phrases, we are able to observe discourse around specific phenomena ("pizzagate," "pandemic," or "Make America Great Again"). In contrast, through more generic culturally significant phrases, we may be able to observe how we organize specific phenomena into the paradigmatic narratives they represent. 70

Much attention has been paid to the use of words and *n*-grams in general in large corpora, but it is difficult to extract from them instances of individual narrative or metaphorical language use. Proverbs, in their tendency to act as both 71 narrative and metaphor, and in their often relatively fixed structure, are an ideal test case for our ability to observe broader cultural narratives through the piecemeal, routine stories employed by the folk. Studies of n-grams tend to organize their interpretation around n-grams which attain a specific historical use-case. In our study of proverbs, we are able to trace the progression of a paradigm and its varying application to historical changes and events.

A natural limitation of this study, and any study that uses extant data to study language, is the issue of representativeness. In this study that limitation is twofold: Both the lexicon for directing the search, and the data being searched are inherently limited. While *The Dictionary of American Proverbs* is extensive, and represents much that is known of proverbs in America, it naturally excludes new proverbs and does not account for many ways in which the structure of the proverbs it contains may be manipulated in their practical use. There are lexical resources that address recent proverbs, for example *The Dictionary of Modern Proverbs*, and the methodology of this study may be readily applied to such lexica [Doyle 2012]. Previous studies on proverb frequency have relied on composite corpora, namely variations of the BNC (British National Corpus), which contains manually curated selections from several domains of text. As shown in this study, composite corpora may miss important differences in proverb usage between domains. 72

Fieldwork (digital and otherwise) continues to be important in identifying new proverbs and changing structures of existing proverbs. This task may be aided in the future by tools like StoryWrangler, that track *n*-gram rank, likely capturing new proverbs in the process. The task then would be extracting likely proverbs from these data, which would require linguistic, cultural, and computational expertise. 73

Much of proverb scholarship has been concerned with the idea of a “paremiological minimum”: A minimum proverbial lexicon for a language and culture. As shown by Lau (1996), and again in the present study, computational studies of the frequency of proverb use can contribute to the understanding of these minima, as those proverbs which seem ubiquitous in large corpora ought to be understood by speakers of a language. Temporal analysis of their frequency may further validate that their frequency is related to enduring currency among the folk, rather than correspondence with a specific occurrence. Another concern in paremiology and phraeseology is the origins of sayings. Work like the present study can serve to both validate and expand on previous scholarship on the history of phrases. 74

In the study of the statistical distribution of natural language, there exists the idea of a kernel lexicon, a subset of words that are essential to communication using a given language. Much literature on the study of culture and education has focused on what one might consider a “minimum of cultural literacy”. Special attention has been paid to which proverbs constitute part of that minimum. It is clear from this study that the most common proverbs vary considerably between corpora. Given the prevalence of these popular proverbs in their respective contexts, we can posit that English learners would benefit in their comprehension of the language if they were familiar with these proverbs. 75

Analyses of the frequency and rank of proverbs in this study (shown in Appendix A) verify that with ever increasing amounts of machine-readable textual data, we may produce longitudinal phraseological studies. 76

Traditionally, the study of metaphor in language has relied on theoretical interpretations [Lakoff 1985]. This study opens up the possibility for new avenues in the study of metaphor that incorporate the currency of specific metaphors among the folk. A natural extension of this study would be employing a similar method to study idioms and conventional metaphors. 77

As machine comprehension of natural language becomes increasingly important, this area too, would benefit from an expanded lexicon that includes proverbs and routine formulae, and understanding of metaphor may be assisted by a more basic understanding of the mapping from general to specific situations that exists in the use of proverbs. 78

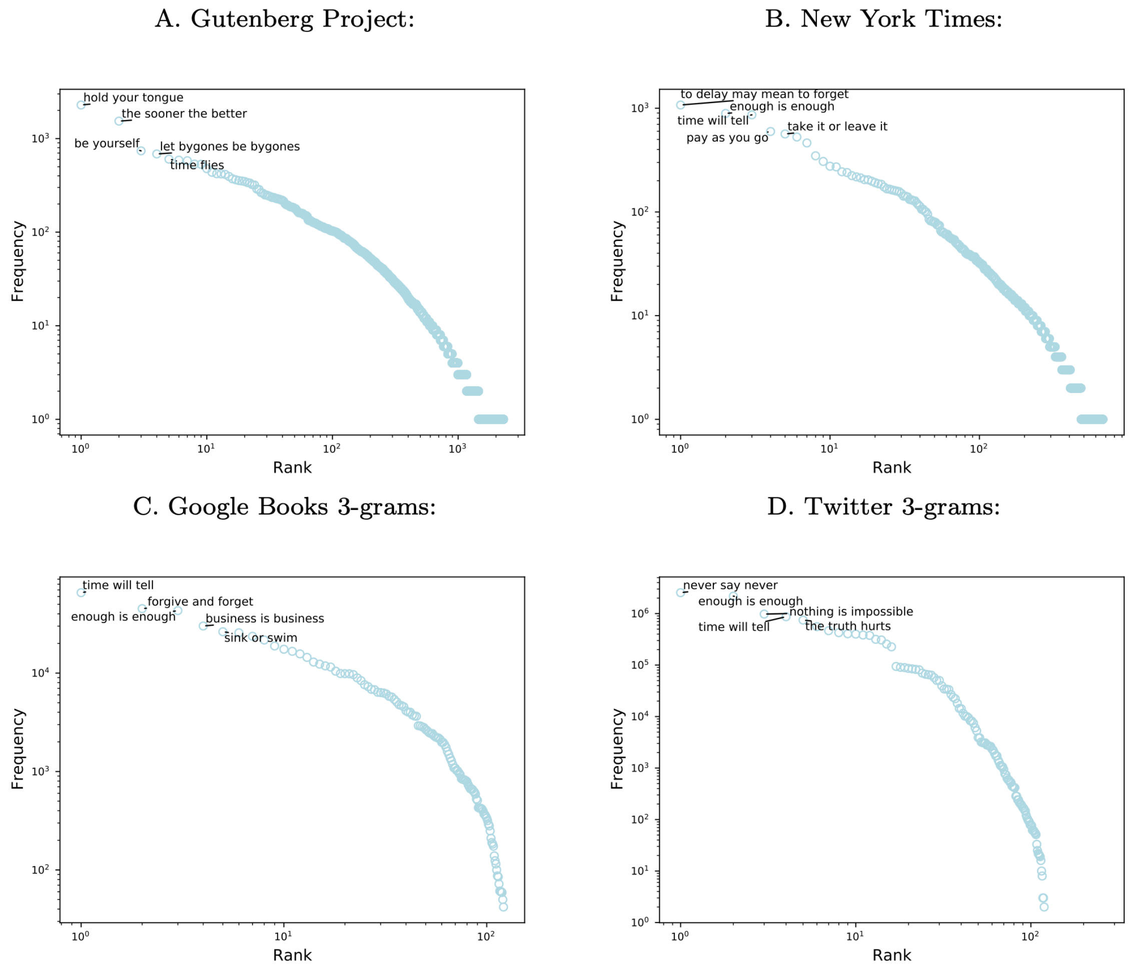
Acknowledgements

Thank you to David Dewhurst, Josh Minot, Michael Arnold, Nicholas Allgaier, Thayer Alshaabi for providing invaluable guidance in the writing of this paper. The authors are grateful for the computing resources provided by the Vermont Advanced Computing Core which was supported in part by NSF award No. OAC-1827314, and financial support from the Massachusetts Mutual Life Insurance Company to CMD and PSD. 79

Appendix A

Figure 9 shows Zipf distributions for entries from Mieder's *Dictionary of American Proverbs* (1992) for each of the four corpora studied, using 3-gram proverbs for Google Books and Twitter. While the distributions exhibit some Zipf’s Law-like behavior (heavy tails), we do not observe robust power-law scaling for all proverbs. We find the largest number of distinct proverbs appearing in Gutenberg and the *New York Times*, on the order of thousands, with the Google Books and Twitter examples showing many fewer. We note that Zipf's law for words does not itself extend over many orders of magnitude [Williams 2015a], typically only 2 or 3, and that it is meaningful, mixed length phrases that present many orders of magnitude of scaling [Williams 2015b]. The Zipf distributions for proverbs are thus comparable to what we see for single words. 80

With a more sophisticated method of proverb detection, one that captures minor variations in phrase structure, we would expect to see some adjustments to the Zipf distributions we have observed, though a priori it is not clear how. Short, robust proverbs (“time flies”) will be well counted, while longer ones for which, say, constituent function words might be changed based on context or era (“he/she/they who hesitates…”) would only see their apparent observed frequency of usage grow. 81



A. Gutenberg Project:

B. New York Times:

C. Google Books 3-grams:

D. Twitter 3-grams:

Figure 9. Zipf distributions for entries from Mieder's *Dictionary of American Proverbs* [Mieder 1992]. For each corpus, proverbs are enumerated and shown on logarithmic axes as a function of rank, with “hold your tongue,” to “delay may mean to forget,” “time will tell,” and “never say never” topping the charts in Gutenberg, NYT, Google, and Twitter respectively. Each distribution exhibits heavy-tailed behavior, more prominently for Gutenberg and NYT.

Appendix B

The data for proverbs in the Gutenberg corpus were used to construct a network with documents as nodes, connected if a given proverb appears in both documents. When betweenness centrality was calculated for nodes in the network, 82 surprisingly James Joyce's *Ulysses* had the 14th highest centrality, close to several dictionaries of proverbs and quotations, and the collected works of Mark Twain (Table B-1). Creasy (2008) documented Joyce's use of proverbs in *Ulysses* from a critical perspective, noting that they are often altered, and blend high and low culture in the work. As Joyce uses many fewer proverbs than a comprehensive proverbial dictionary, the book's centrality in this network implies that Joyce's use of proverbs is far from arbitrary, and that his choice of proverbs is purposefully situated in the broader context of English proverbial knowledge.

	Book	**btwn centrality**
1	Dictionary of Quotations	0.043048
2	Familiar Quotations	0.022821
3	Dictionary of English Proverbs and Proverbial Phrases	0.014274
4	A Polyglot of Foreign Proverbs	0.013061
5	The Entire Project Gutenberg Works of Mark Twain	0.013041
6	French Idioms and Proverbs	0.010083
7	Roget's Thesaurus	0.009785
8	Webster's Unabridged Dictionary	0.007978
9	U.S. Copyright Renewals 1950 – 1977	0.006709
10	The Project Gutenberg Complete Works of Gilbert Parker	0.006278
11	Proverb Lore	0.006028
12	Complete Project Gutenberg John Galsworthy Works	0.003897
13	Complete Project Gutenberg Works of George Meredith	0.003660
14	Ulysses	0.003184
15	The Historical Romances of Georg Ebers	0.003168
16	Familiar Quotations	0.003007
17	The Circle of Knowledge	0.002886
18	The Complete Poetic and Dramatic Works of Robert Browning	0.002749
19	Complete Project Gutenberg Oliver Wendell Holmes, Sr. Works	0.002657
20	Motion Pictures, 1960-1969: Catalog of Copyright Entries	0.002578

Table 1. The 20 most central books by betweenness centrality, from a network of books connected by shared proverbs in Gutenberg. Notably, James Joyce's *Ulysses* appears alongside several proverb and quotations collections, and the collected works of Mark Twain.

Appendix C

Tables 2-4 show the total count of the 50 most popular proverbs in 83

their respective corpora. 84

	Proverb	**Count**
1	hold your tongue	2,284
2	the sooner the better	1,536
3	be yourself	739
4	let bygones be bygones	685
5	time flies	603
6	alls well that ends well	588
7	one thing at a time	580

8	business is business	534
9	sink or swim	531
10	forgive and forget	477
11	take it or leave it	436
12	nothing is impossible	421
13	better late than never	419
14	every man for himself	414
15	know thyself	394
16	share and share alike	372
17	slow but sure	363
18	live and let live	356
19	the more the merrier	352
20	the die is cast	348
21	honesty is the best policy	339
22	to be or not to be	335
23	do or die	322
24	never say die	319
25	extremes meet	289
26	art for arts sake	286
27	all men are created equal	265
28	let well enough alone	260
29	time is money	250
30	no accounting for taste	249
31	peace at any price	244
32	tastes differ	241
33	history repeats itself	235
34	boys will be boys	235
35	charity begins at home	231
36	love is blind	228
37	the end justifies the means	227
38	one good turn deserves another	224
39	blood is thicker than water	221
40	not wisely but too well	219
41	all things work together for good	213
42	first come first served	201
43	keep the wolf from the door	196
44	dead men tell no tales	195
45	the wages of sin is death	191
46	seeing is believing	187
47	keep a stiff upper lip	186
48	ignorance is bliss	185
49	where theres a will theres a way	183
50	murder will out	179

Table 2. The top 50 proverbs and proverbial expressions (from the *Dictionary of American Proverbs*) in the entire Gutenberg Corpus.

	Proverb	Count
1	to delay may mean to forget	1,075
2	enough is enough	891
3	time will tell	864
4	pay as you go	597
5	take it or leave it	565
6	do or die	528
7	first come first served	463
8	be yourself	348
9	father knows best	307
10	never say never	276
11	live and let live	272
12	money talks	244
13	the sooner the better	240
14	better late than never	224
15	sink or swim	218
16	boys will be boys	213
17	time flies	205
18	time is of the essence	204
19	divide and conquer	198
20	gentlemen prefer blondes	192
21	to be or not to be	187
22	the show must go on	185
23	time is money	174
24	talk is cheap	167
25	every man for himself	166
26	leave well enough alone	163
27	put up or shut up	161
28	business is business	159
29	accentuate the positive	157
30	forgive and forget	151
31	you get what you pay for	142
32	safety first	142
33	too little and too late	140
34	there is no easy way	132
35	let the chips fall where they may	131
36	all men are created equal	129
37	the more the merrier	128
38	history repeats itself	122
39	let bygones be bygones	117
40	one thing at a time	113

41	let nature take its course	106
42	never say die	106
43	seeing is believing	102
44	nothing is impossible	100
45	war is hell	95
46	the worst is yet to come	85
47	actions speak louder than words	82
48	gone but not forgotten	82
49	to each his own	80
50	let the buyer beware	80

Table 3. The top 50 proverbs and proverbial expressions (from the *Dictionary of American Proverbs*) in the *New York Times* from 1987-2007.

	Proverb	**Count**
1	hold your tongue	131,426
2	time will tell	65,640
3	forgive and forget	45,189
4	enough is enough	43,149
5	business is business	30,101
6	sink or swim	26,315
7	nothing is impossible	25,695
8	easy does it	23,655
9	do or die	21,672
10	time is money	18,856
11	practice makes perfect	17,469
12	never say never	16,649
13	divide and conquer	15,673
14	love is blind	14,439
15	seeing is believing	12,951
16	never say die	12,329
17	ignorance is bliss	11,838
18	history repeats itself	11,529
19	fair is fair	10,456
20	slow but sure	9,898
21	forewarned is forearmed	9,860
22	love conquers all	9,839
23	misery loves company	9,654
24	facts are facts	8,944
25	time will pass	8,389
26	orders are orders	7,620
27	the truth hurts	7,292
28	blood will tell	6,840
29	father knows best	6,783

30	try anything once	6,388
31	murder will out	6,349
32	silence is golden	6,278
33	war is hell	6,136
34	business before pleasure	5,811
35	talk is cheap	5,723
36	revenge is sweet	5,400
37	familiarity breeds contempt	5,095
38	might makes right	4,768
39	consider the source	4,677
40	toe the mark	4,549
41	every little helps	4,139
42	time marches on	4,019
43	nothing is perfect	4,007
44	money is power	3,757
45	circumstances alter cases	3,668
46	respect your elders	3,644
47	gentlemen prefer blondes	2,922
48	mother knows best	2,908
49	love never fails	2,848
50	nobody is perfect	2,801

Table 4. The top 50 3-gram proverbs and proverbial expressions (from the *Dictionary of American Proverbs*) in the Google Books Ngram Corpus.

	Proverb	**Count**
1	never say never	2,549,095
2	enough is enough	2,182,460
3	nothing is impossible	978,533
4	time will tell	869,662
5	the truth hurts	748,285
6	forgive and forget	557,294
7	talk is cheap	465,608
8	love is blind	426,010
9	practice makes perfect	405,635
10	nobody is perfect	399,324
11	time is money	383,632
12	ignorance is bliss	377,037
13	do or die	316,328
14	history repeats itself	307,467
15	love never fails	255,795
16	misery loves company	226,217
17	divide and conquer	94,085
18	facts are facts	90,513

19	respect your elders	89,372
20	seeing is believing	86,169
21	time will pass	84,432
22	silence is golden	82,346
23	love conquers all	80,964
24	revenge is sweet	69,820
25	health is wealth	66,274
26	never say die	65,115
27	prayer changes things	63,757
28	iron sharpens iron	57,065
29	sink or swim	50,361
30	tomorrow never comes	50,297
31	business is business	39,525
32	hold your tongue	34,344
33	nothing is perfect	34,050
34	try anything once	33,370
35	mother knows best	26,848
36	every little helps	23,672
37	never waste time	22,244
38	fair is fair	18,125
39	slow but sure	14,404
40	consider the source	14,201
41	justice is blind	11,604
42	money is power	10,186
43	time works wonders	10,079
44	time changes everything	9,512
45	like attracts like	8,320
46	familiarity breeds contempt	8,166
47	war is hell	7,439
48	easy does it	6,071
49	gentlemen prefer blondes	5,273
50	courtesy costs nothing	3,890

Table 5. The top 50 proverbs and proverbial expressions (from the *Dictionary of American Proverbs*) on Twitter from 2008-2021.

Works Cited

Abello 2012 Abello, J., Broadwell, P. and Tangherlini, T. R. (2012a) “Computational folkloristics,” *Communications of the ACM*, 55(7), pp. 60–70. doi: 10.1145/2209249.2209267.

Abrams 1994 Abrams, R. D. and Babcock, B. A. (1994) “The literary use of proverbs,” in Mieder, W. (ed.) *Wise words: Essays on the Proverb*. New York: Garland (Garland reference library of the humanities, vol. 1638), pp. 415–437.

Albig 1931 Albig, W. (1931) “Proverbs and social control,” *Sociology and Social Research*, (15), pp. 527–535.

Alshaabi 2020 Alshaabi, T. *et al.* (2020) “Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter,” *arXiv:2007.12988 [physics]*. Available at: http://arxiv.org/abs/2007.12988 (Accessed: 7 December 2020).

Alter 2018 Alter, C. (2018) "The Young and the Relentless," *TIME*. Available at: https://time.com/magazine/us/5210502/april-2nd-2018-vol-191-no-12-u-s/.

Andersson 2013 Andersson, D. (2013) "Understanding figurative proverbs: A model based on conceptual blending," *Folklore*, 124(1), pp. 28–44. doi: 10.1080/0015587X.2012.734442.

Bain 1939 Bain, R. (1939) "Verbal stereotypes and social control," *Sociology and Social Research*, (23), pp. 431–446.

Balasubrahmanyan 1996 Balasubrahmanyan, V. K. and Naranan, S. (1996) "Quantitative linguistics and complex system studies," *Journal of Quantitative Linguistics*, 3(3), pp. 177–228. doi: 10.1080/09296179608599629.

British National Corpus 2001 "British National Corpus," (2001). Available at: http://www.natcorp.ox.ac.uk.

Cancho 2001 Cancho, R. F. i. and Solé, R. V. (2001) "Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited," *Journal of Quantitative Linguistics*, 8(3), pp. 165–173.

Cancho 2003 Cancho, R. F. i. and Sole, R. V. (2003) "Least effort and the origins of scaling in human language," *Proceedings of the National Academy of Sciences*, 100(3), pp. 788–791. doi: 10.1073/pnas.0335980100.

Clancy 1990 Clancy, T. (1990). *Clear and present danger*. Berkley mass-market edition. ed. Berkley Books, New York.

Clauset 2009 Clauset, A., Shalizi, C. R. and Newman, M. E. J. (2009) "Power-law distributions in empirical data," *SIAM Review*, 51(4), pp. 661–703. doi: 10.1137/070710111.

Corominas-Murtra 2010 Corominas-Murtra, B. and Solé, R. V. (2010) "Universality of Zipf's law," *Physical Review E*, 82(1), p. 011102. doi: 10.1103/PhysRevE.82.011102.

Corral et al. 2015 Corral, A., Boleda, G. and Ferrer-i-Cancho, R. (2015) "Zipf's law for word frequencies: Word forms versus lemmas in long texts," *PLOS ONE*, 10(7), p. e0129031. doi: 10.1371/journal.pone.0129031.

Cram 1994 Cram, D. (1994) "The linguistic status of the Proverb," in Mieder, W. (ed.) *Wise Words: Essays on the Proverb*. New York: Garland (Garland reference library of the humanities), pp. 73–97.

Creasy 2008 Creasy, M. (2008) "'To vary the timehonoured adage': Ulysses and the Proverb," *English*, 57(217), pp. 65–81. doi: 10.1093/english/efn008.

Dodds 2017 Dodds, P. S. et al. (2017) "Simon's fundamental rich-get-richer model entails a dominant first-mover advantage," *Physical Review E*, 95(5), p. 052301. doi: 10.1103/PhysRevE.95.052301.

Dodds 2021 Dodds PS, Minot JR, Arnold MV, Alshaabi T, Adams JL, Reagan AJ, et al. (2021) "Computational timeline reconstruction of the stories surrounding Trump: Story turbulence, narrative control, and collective chronopathy," PLoS ONE 16(12): e0260592. https://doi.org/10.1371/journal.pone.0260592

Doyle 2012 Doyle, C. C., Mieder, W. and Shapiro, F. R. (2012) *The Dictionary of Modern Proverbs*. New Haven: Yale University Press.

Dundes 1965 Dundes, A. (1965) "The study of folklore in literature and culture: Identification and interpretation," *The Journal of American Folklore*, 78(308), p. 136. doi: 10.2307/538280.

Fazly 2009 Fazly, A., Cook, P. and Stevenson, S. (2009) "Unsupervised Type and Token Identification of Idiomatic Expressions," *Computational Linguistics*, 35(1), pp. 61–103. doi: 10.1162/coli.08-010-R1-07-048.

Georgia 1991 Georgia, S. C. of (1991) "Cunningham v. State 1991."

Glaberson 1992 Glaberson, W. (1992) "Assault case renews debate on rape shield law," *The New York Times*.

Haas 2008 Haas, H. A. (2008) "Proverb familiarity in the United States: Cross-regional comparisons of the paremiological minimum," *Journal of American Folklore*, 121(481), pp. 319–347. doi: 10.2307/20487611.

Haas 2022 Haas, H. A. (2022) "The Proverbs of a Pandemic: The Early Months of the COVID-19 Pandemic Viewed through the Lens of Google Trends," *Journal of American Folklore*, 135 (535), pp. 26–48. Doi: https://doi.org/10.5406/15351882.135.535.02

Hanley 1992 [Hanley, R. (1992) "Jury chosen in Glen Ridge assault trial," *The New York Times*.

Henry 1996 Henry, J. (1996) "How the money was spent in previous environmental Bond Acts," *The New York Times*.

Hunter 2007 Hunter, J. D. (2007) "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, 9(3), pp.

90–95. doi: 10.1109/MCSE.2007.55.

Koplenig 2018 Koplenig, A. (2018) “Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes — a large-scale corpus analysis,” *Corpus Linguistics and Linguistic Theory*, 14(1), pp. 1–34. doi: 10.1515/cllt-2014-0049.

Lakoff 1985 Lakoff, G. and Johnson, M. (1985) *Metaphors We Live By*. Chicago, Ill.: Univ. of Chicago Press.

Lau 1996 Lau, K. J. (1996) “'It’s About Time': The ten proverbs most frequently used newspapers and their relation to American values,” *Proverbium*, 13, pp. 135–59.

Leskovec 2009 Leskovec, J., Backstrom, L. and Kleinberg, J. (2009) “Meme-tracking and the dynamics of the news cycle,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’09*. Paris, France: ACM Press, p. 497. doi: 10.1145/1557019.1557077.

Lieber 1994 Lieber, M. D. (1994) “Analogic ambiguity: A paradox of proverb usage,” in Mieder, W. (ed.) *Wise Words: Essays on the Proverb*. New York: Garland (Garland reference library of the humanities, vol. 1638), pp. 99–126.

Love is Blind 2020 “Love Is Blind” (2020). Netflix.

Mandelbrot 1953 Mandelbrot, B. (1953) “An informational theory of the statistical structure of languages,” in Jackson, W. (ed.) *Communication Theory*. Academic Press, Princeton, pp. 486–502.

March for Our Lives 2018 “March for Our Lives Highlights: Students Protesting Guns Say 'Enough Is Enough'” (2018) *The New York Times*.

Martínez-Mekler 2009 Martínez-Mekler, G. *et al.* (2009) “Universality of rank-ordering distributions in the arts and sciences,” *PLoS ONE*. Edited by M. Costa, 4(3), p. e4791. doi: 10.1371/journal.pone.0004791.

Mechling 2004 Mechling, J. (2004) ““Cheaters Never Prosper” and other lies adults tell kids: Proverbs and the culture wars over character,” in Lau, K. J., Tokofsky, P., and Winick, S. D. (eds) *What goes around comes around: the circulation of proverbs in contemporary life: Essays in Honor of Wolfgang Mieder*. Logan: Utah State University Press, pp. 107–126.

Michel 2011 Michel, J.-B. *et al.* (2011) “Quantitative analysis of culture using millions of digitized books,” *Science*, 331(6014), pp. 176–182. doi: 10.1126/science.1199644.

Mieder 1992 Mieder, W., Kingsbury, S. A. and Harder, K. B. (eds) (1992) *A Dictionary of American Proverbs*. New York: Oxford University Press.

Mieder 2004 Mieder, W. (2004) *Proverbs: a handbook*. Westport, Conn: Greenwood Press (Greenwood folklore handbooks).

Mieder 2005 Mieder, W. (2005) *Proverbs are the best policy: Folk wisdom and American politics*. Logan, Utah: Utah State University Press.

Mieder 2008 Mieder, W. (2008) *'Proverbs speak louder than words': Folk wisdom in art, culture, folklore, history, literature and mass media*. New York: P. Lang.

Mieder 2012 Mieder, W. (2012) *Proverbs are never out of season: Popular wisdom in the modern age*. New York: Peter Lang (International folkloristics, v. 7).

Mieder 2019 Mieder, W. (2019) *“Right makes Might”: Proverbs and the American worldview*. Bloomington, Indiana: Indiana University Press.

Moon 1998 Moon, R. (1998) *Fixed expressions and idioms in English: A corpus-based approach*. Oxford : New York: Clarendon Press ; Oxford University Press (Oxford studies in lexicography and lexicology).

Norrick 1985 Norrick, N. R. (1985) *How Proverbs Mean: Semantic Studies in English Proverbs*. Berlin, New York: DE GRUYTER MOUTON. doi: 10.1515/9783110881974.

Obelkevich 1994 Obelkevich, J. (1994) “Proverbs and social history,” in Mieder, W. (ed.) *Wise Words: Essays on the Proverb*. New York: Garland (Garland reference library of the humanities, vol. 1638), pp. 211–252.

Pechenick 2015a Pechenick, Eitan Adam, Danforth, C. M. and Dodds, P. S. (2015) “Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution,” *PLOS ONE*. Edited by A. Barrat, 10(10), p. e0137041. doi: 10.1371/journal.pone.0137041.

Pechenick 2015b Pechenick, Eitan A., Danforth, C. M. and Dodds, P. S. (2015) “Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not.”

Permiakov 1989 Permiakov, G. L. (1989) “On the question of a Russian paremiological minimum.,” *Proverbium*, 6, pp. 91–102.

Reagan 2016 Reagan, A. J. *et al.* (2016) “The emotional arcs of stories are dominated by six basic shapes,” *EPJ Data Science*, 5(1), p. 31. doi: 10.1140/epjds/s13688-016-0093-1.

Robinson 202 Robinson, D. (2020) *Gutenbergr*. Available at: https://cran.r-project.org/web/packages/gutenbergr/index.html.

Sag 2002 Sag, I. A. *et al.* (2002) “Multiword expressions: A pain in the neck for NLP,” in Goos, G. et al. (eds) *Computational Linguistics and Intelligent Text Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–15. doi: 10.1007/3-540-45715-1_1.

Sandhaus 2008 Sandhaus, E. (2008) “The New York Times Annotated Corpus,” *Linguistic Data Consortium*. doi: 10.35111/77BA-9X74.

Semones 202 Semones, E. (2020) “'Enough is enough': Thousands descend on D.C. for largest George Floyd protest yet,” *POLITICO*.

Shutova 2010 Shutova, E. (2010) “Models of metaphor in NLP,” in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 688–697. Available at: https://www.aclweb.org/anthology/P10-1071.

Simon 1960 Simon, H. A. (1960) “Some further notes on a class of skew distribution functions,” *Information and Control*, 3(1), pp. 80–88. doi: 10.1016/S0019-9958(60)90302-8.

Steyer 2015 Steyer, K. (2015) “Proverbs from a corpus linguistic point of view,” in Hrisztova-Gotthardt, H. and Aleksa Varga, M. (eds) *Introduction to Paremiology: A Comprehensive Guide to Proverb Studies*. Warsaw, Poland: DE GRUYTER OPEN, pp. 206–228. doi: 10.2478/9783110410167.

Swuste 2010 Swuste, P., Gulijk, C. van and Zwaard, W. (2010) “Safety metaphors and theories, a review of the occupational safety literature of the US, UK and The Netherlands, till the first part of the 20th century,” *Safety Science*, 48(8), pp. 1000–1018. doi: 10.1016/j.ssci.2010.01.020.

Tangherlini 2016 Tangherlini, T. R. (2016) “Big folklore: A special issue on computational folkloristics,” *The Journal of American Folklore*, 129(511), p. 5. doi: 10.5406/jamerfolk.129.511.0005.

Tweedle 2012 Tweedle, V. and Smith, R. J. (2012) “A mathematical model of Bieber Fever: The most infectious disease of our time?,” in.

Underwood 2018 Underwood, T., Bamman, D. and Lee, S. (2018) “The transformation of gender in English-language fiction,” p. 25.

Vote Yes on the Bond Act 1996 “Vote Yes on the Bond Act,” (1996) *The New York Times*.

Whiting 2014 Whiting, B. J. (2014) *Modern Proverbs and Proverbial Sayings*. Available at: https://0-doi-org.pugwash.lib.warwick.ac.uk/10.4159/harvard.9780674864153.

Williams 2015a Williams, J. R. *et al.* (2015a) “Text mixing shapes the anatomy of rank-frequency distributions,” *Physical Review E*, 91(5), p. 052811. doi: 10.1103/PhysRevE.91.052811.

Williams 2015b Williams, J. R. et al. (2015b) “Zipf’s law holds for phrases, not words,” *Scientific Reports*, 5(1), p. 12209. doi: 10.1038/srep12209.

Zemeckis 1994 Zemeckis, R. (1994) *Forrest Gump*. Paramount Pictures.

Zipf 2012 Zipf, G. K. (2012) *Human behavior and the principle of least effort: An introduction to human ecology.* Mansfield Centre, Conn: Martino Publishing [u.a.].

pandas 2020 The pandas development team (2020) *pandas-dev/pandas: Pandas*. Zenodo. doi: 10.5281/zenodo.3509134.

Özbal 2016a Özbal, G. *et al.* (2016) “Learning to identify metaphors from a corpus of proverbs,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2060–2065. doi: 10.18653/v1/D16-1220.

Özbal 2016b Özbal, G., Strapparava, C. and Sinem Tekiroglu, S. (2016) “PROMETHEUS: A corpus of proverbs annotated with metaphors,” *LREC, Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC’16), pp. 3787–3793.

Čermák 2014 Čermák, F. (2014) *Proverbs: their lexical and semantic features*. Burlington, Vermont: The University of Vermont (Supplement series of Proverbium Yearbook of International Proverb Scholarship, volume 36).