A Short Primer on Historical Natural Language Processing

Thomas Hills*and Alessandro Miani†

“The past is a foreign country: they do things differently there.” LP Hartley

Natural language processing and large databases of digitized text have made it trivial to compute statistics over millions of words of unstructured language data. One of the goals of this kind of work is making inferences about the psychology of authors. Are they happy? Are they depressed? Are they showing signs of late life cognitive decline? What is the state of their mental health? What do they think about the new Star Wars movie? What do they think about gender roles, the future, or the economy? What kinds of things are they talking about? What do they believe and why? Natural language processing combined with behavioural insights is allowing us to answer questions like these at unprecedented scales.

When these kinds of questions are focused on longitudinal data, researchers are able to ask questions about how behaviour, psychology, and culture changes over time. How does language evolve? Is language becoming more abstract, more information rich, or easier to understand? Were people happier or sadder in the past? What factors influenced these changes? How has attention to rationality or mental health changed over time? What about within an individual? How did Darwin’s views change across his lifetime? Do the letters of Mozart show a relationship between mental suffering and artistic output? Do the books or speeches of prolific language producers signal growing vulnerability to cognitive decline?

The goal of this chapter is help readers see how this work is done, the various methodologies employed in doing it, the kinds of resources available, and the potential pitfalls and ways to avoid them. We will also provide and overview of some of the exceptional work in this area.

Historical Natural Language Processing

There are numerous excellent accounts of historical psychology (e.g., Pinker 2011; Bourke 2015). Those accounts are often focused on the psychology of individuals who experienced that history based on their personal accounts. This research has traditionally involved a *close read* of texts from original sources. Researchers make inferences by reading the original sources *with their eyes.* In some cases, these inferences are supported with additional data, such as numbers of people dying from certain events or changes in economic output. The close read is usually considered qualitative, as it focuses on the many nuanced qualities of the writing. The benefit is that one can sensitively detect much of the meaning intentionally conveyed by the author. The cost is that this sensitive interpretation is more or less subjective, it is difficult and costly to replicate, and it is limited to what the reader can reasonably read and remember.

*University of Warwick, t.t.hills@warwick.ac.uk

†University of Neuchâtel

Digitized databases of historical language data have given rise to a different approach: the *distant read* involving computer-assisted quantification of texts, or what many people simply call *natural language processing* (NLP). More precisely, NLP is quantitative and involves algorithms that analyse historical language data through large scale statistical analysis. Eyeballs are optional.

This work is 'distant' for several reasons. It typically involves analyzing more documents than any boundedly rational being could read or remember. It is often unconcerned with nuanced interpretation of documents; rarely does it focus on anything as small as a single sentence, paragraph, or chapter. Indeed, this approach often treats a text as a *bag of words*, with grammatical information and word order swept under the rug. In practice, this approach is capable of summarizing vast amounts of language data, covering millions of writers over the course of hundreds of years of history. The cost of this work is that it lacks this nuanced interpretation of the close read. The benefit is that it can capture large-scale trends, it can process more data than any individual can read, and it is fast, replicable, and it can use all of the available data, which can wash out sampling error. What the data loses in detailed qualitative analysis, it gains in data quantity and quantitative rigour.

Historical NLP aiming to infer psychology varies in complexity, but the majority of this work can be classified into three distinct approaches: 1) Counting words or documents, 2) Averaging the semantic meanings of words along some dimension (e.g., the positive or negative valence of words used in sentiment analysis), or 3) Sorting, for example, using machine learning methods such as topic modelling. The goal with sorting is to organize and identify text around specific dimensions so that inferences can be made about the qualities of these dimensions. These three methods provide a breadth and flexibility that allow researchers to address a wide array of questions, as the examples below demonstrate.

Word counting

The majority of historical language analysis has focused on counting words and computing their proportion or relative frequency in the text. This can be as simple as computing the relative frequency of individual words to detect their rise and fall over historical time. Michel et al. (2011) article on *culturomics* was one of the first to take this approach from the perspective of big data and also—because it was associated with Google—presented the first large scale analysis of the Google Ngram Corpus.[1]

The Google Ngram Corpus is a collection of digitized books in various languages (e.g., English, British English, American English, Chinese, Russian, etc), and from a variety of sources (e.g., fiction or non-fiction books), representing a small but substantial fraction of all the books published over the last several hundred years. The data becomes fairly sparse if one goes back beyond 1800, but after this time users can investigate the frequencies of various words and phrases at sufficient scale to address many different kinds of problems.

The primary focus of Michel et al. (2011) was to demonstrate the possibilities of the corpus and most of the examples involved word counting. For example, how has the frequency of regular and irregular verbs changed over the last few hundred years in American and British

[1] Though other historical language corpora existed at the time, such as the Corpus of Historical American English (Davies 2009).

English? Reading individual books to detect the quantitative change over the course of several hundred years would be an unsavory task. But with a large historical corpus, the problem becomes somewhat trivial: by counting word frequencies one can, for example, see that it took more than 200 years for a word like “chide” to replace its irregular forerunner (“chode”). The frequency of most of the irregular verbs (Michel et al. 2011) present follows this pattern (e.g., sped, chode, burnt). One can go to the Google Ngram Viewer (https://books.google.com/ngrams) and test out words of one’s own.

The frequencies of well chosen words can reveal enticing patterns. Greenfield (2013) used the Google Ngram corpus to examine theories about urban and rural social change by examining the frequency of words like “choose” vs “obliged” and “give” vs “get”, which roughly followed trends in rising urban and declining rural populations. By choosing theory-motivated words, Greenfield was able to provide evidence for the hypothesis that transitions to more urban environments correspond to increasing individualistic and materialistic values.

Though methodologically trivial, this work still has a great deal of promise and is probably largely untapped. Recent high-profile work takes a very similar approach using a large publically available British newspaper corpus (Lansdall-Welfare et al. 2017).

Dictionaries

A more sophisticated approach to word counting is to count groups of words. For example, one can ask how do words or phrases related to the self, the future or the past (e.g., dates), cognitive distortions, cognitive tension, and so on, increase or decrease over time. With this approach, one defines collections of words that are associated with particular patterns of thought and computes their relative frequencies.

A useful and automated tool for this kind of analysis is Linguistic Inquiry and Word Count (LIWC, https://www.liwc.app/, Boyd et al. 2022; Pennebaker et al. 2022), which contains numerous pre-defined dictionaries (translated into multiple languages) related to basic cognitive frames such politics, present focus, insight, and social behavior. Using LIWC, one can rapidly compute the relative frequency of words associated with these dictionaries and compare texts over time, across different experimental conditions, and from different groups of individuals.

Boyd, Blackburn, and Pennebaker (2020) recently used this approach to examine more than 50,000 narratives from publically available corpora to demonstrate that narrative arcs tend to follow standard patterns of words related to ‘staging’, ‘plot progression’, and ‘cognitive tension’. This was achieved by breaking documents up into proportional segments and then computing the relative frequency for each dictionary over the segments.

LIWC is extremely rich and can measure everything from pronoun frequencies to cognitive tension. The risk with LIWC is that running all of these measures to detect differences over time or between populations will likely turn up something, just by chance. This may tempt one to hypothesize after the results are known (HARKing)—to make up a story to explain why a particular dictionary shows the pattern it does. However, LIWC can be a valuable exploratory tool and with proper discipline and transparency can be a valuable research tool. The best work using LIWC often uses strong theory and multiple corpora to validate findings across numerous contrasts. See for example (MacKrill et al. 2021), looking at the language properties correlated with TED talk views and emotional responses, and (Sap et al. 2022) comparing autobiographical versus imagined content.

Developing ad hoc dictionaries. If one has a particular research question, one can develop a dictionary to address it. Bollen et al. (2021) used a dictionary derived from experts in Cognitive Behavioral Therapy to identify phrases associated with what they call *cognitive distortion schemata*. These represent things like mind-reading ("everyone thinks"), emotional reasoning ("still feels"), and labeling ("I am a"). By examining the relative frequencies of these schemata over time, Bollen and colleagues found that these distortions have risen dramatically in the last 50 years. Scheffer et al. (2021) used an unsupervised machine learning approach (Principal Components Analysis) to artificially construct dictionaries and then, using this exploratory approach, focused in on specific word types to examine historical trends in rationality.

One can also come up with clever and elegant kinds of dictionaries, such as dates, to examine prospective versus retrospective thinking. Preis et al. (2012) developed a *future orientation index* based on the relative frequencies of future and past dates in internet searches. Using a cross-national sample, they found that countries with more prospective search also had a higher Gross Domestic Product. Müller and Schwarz (2021) looked at frequencies of anti-immigrant speech on social media and found a positive correlation with hate crime across different municipalities in Germany.

Fast, Chen, and Bernstein (2016) developed an open-source LIWC-like Python package called *Empath* (see also Fast, Chen, and Bernstein 2017). Empath is provided with 194 built-in, pre-validated content dictionaries (e.g., *deception*, *health*, *money*, *religion*, etc) that highly correlate with LIWC's (average correlation of $r = .91$ ranging from .86, against LIWC's *work* dictionary, to .94, against *positive emotion*). Besides being freely available, what makes Empath different from LIWC (whose construction relied on human coding) is that Empath dictionaries have been generated in an unsupervised fashion relying on a set of three diverse corpora: 1) The New York Times articles from 1987 to 2007, 2) posts from the social media Reddit from 2008 to 2015, and 3) amateur fiction writings from wattpad.com.

Empath is built via neural embedding using the skip-gram architecture of the *word2vec* algorithm (Mikolov et al. 2013). From each corpus, the authors generated a space $M(n \times h)$ where n is the size of vocabulary, and h the number of hidden nodes in the network. In this space, each vector v is a word (or ngram, i.e., a concatenation of n words) whose dimensions defines the weight of its connection to one of the hidden layer neurons h. The authors computed the cosine similarity between word vectors and then extracted, given a seed word, the most similar words to build a specific dictionary. This reasoning is based on the fact that similar (i.e., semantically close) words share similar contexts (e.g., *doctor* and *nurse* are closer and share more context than *doctor* and *bread*).

Although it seems, at the time of writing, that the Empath project is discontinued (the authors provided a web service to build *ad hoc* dictionaries, http://empath.stanford.edu, now vanished), the legacy left by Empath is invaluable. Via the Python package, Empath allows one to create *ad hoc* dictionaries. Few steps are required: define the corpus model from which to build the dictionary (one of the three provided: New York Times, Reddit, and Fiction); define a set of word seeds around which the dictionary will be built (e.g., for a dictionary named *colors*, a set of seeds would be: [“red”,“blue”,“green”]); define the size of the dictionary in terms of words returned. Researchers interested in the Cold War, for example, could build a dictionary from the New York Times articles using the seed “cold_war”, obtaining a set of N words (w_N) order by similarity: w_1 = *cold_war*; w_2 = *the_cold_war*; w_3 = *the_Cold_War*; w_{10} = *Soviet_power*; w_{50} = *postwar_period*; w_{100} = *fascism*; w_{150} = *foreign_policy*; w_{180} = *political_struggle*. Note that the words returned have to be pre-processed according to the corpus they are tested against in order to match tokens (words or ngrams). In fact, if there is no pre-processing match between the dictionary and the corpus (e.g., character cases, stopwords, ngramming), the dictionary is less effective in extrapolating the dictionary from the text.

Crucial for historical NLP, Empath allows one to create LIWC-like categories that are historically sound. Because words change meaning during time (Li et al. 2019; Li, Hills, and Hertwig 2020), a semantic historical mismatch between categories and the corpus could undermine results (although methods have been developed for correcting such issue, see Thomas T. Hills et al. 2019). To overcome this problem, dictionaries could be compiled from space models built on historical corpora. The Google Ngram Corpus (Michel et al. 2011) represents a precious resource for this endeavor, offering enough material to train space models for each decade or year, at least from 1800. Space models based on *word2vec* architecture could be easily built with 10-20 lines of code via R (e.g., *text2vec*) or Python (e.g., *gensim*) packages from any corpus available to, or constructed by, researchers.

Besides LIWC and Empath, there are several dictionaries available Bollen et al. (2021), and often, the authors provide their dictionaries as supplemental data. The R package *qdapDictionaries* provides a series of freely available dictionaries such as lists of words related to power, strength, weakness, but also negation (e.g., “not”, “never”) and function words (e.g., “the” and “of”) as well as the top 25, 100, 200, and 1000 most frequent English words. Once a dictionary is built, or collected, then, applying it to texts can be achieved via the R package *quanteda* or the Python *Empath*. Note that LIWC allows researchers to use *ad hoc* dictionaries, and in fact, LIWC offers a list of user-generated dictionaries accessible upon purchase of the licence code. For the less NLP-experienced researchers, the Custom List Analyzer (CLA, Kyle, Crossley, and Kim 2015) provides an *ad hoc* list of dictionaries on batches of texts via a standalone application.

Word Distributions

Some additional counting methods are also worth noting. For example, one can compute the distribution of word frequencies. The most notable example of this is *Zipf’s law*: word frequencies in texts plotted on a log-log plot–with frequency on the y-axis and frequency-rank on the x-axis–frequently produce a straight line over several orders of magnitude (Zipf 1949). Zipf observed that the slope of this line was approximately -1.

The characteristics of Zipf’s distribution are different for different texts. The slope is also changing over historical time. Using the same frequency data, one can also compute the *entropy*: how predictable is the language as a whole. Pilgrim, Guo, and Hills (2021) recently found a rising trend in entropy over the last several hundred years, suggesting that Zipf distributions may be flattening in more recent years, with the language as a whole becoming less predictable.

Type-token ratio (TTR) is another useful indicator. It computes the number of word types (different kinds of words) over the number of tokens (the total number of words including repetitions of the same type). It provides a measure of the sophistication or diversity of the language: high values indicate high rates of unique words (where 1 means that no words are repeated). (Le et al. 2011) used this, among other measures, as a potential indicator of age-related cognitive decline among three different authors (Agatha Christie, Iris Murdoch, and P.D. James) who published numerous books over the course of their lives. The question they explored was whether or not Agatha Christie had Alzheimer's. She was suspected of having it. Iris Murdoch died with Alzheimer's and P.D. James aged healthily. The results are provocative. More recent work has expanded this to mental health using many different indicators combined with machine learning (e.g., Orimaye et al. 2017; Fraser, Meltzer, and Rudzicz 2016).

The R packages *koRpus* and *quanteda* compute the TTR offering different formulas, attempting to overcome the problem that TTR is non-linearly related to document length (i.e., the longer the text, the lower the TTR). TAALED (Tool for the Automatic Analysis of LExical Diversity, Zenker and Kyle 2021) is a standalone application designed to calculate a wide variety of lexical diversity indices based on TTR. Interestingly, TAALED returns the TTR computed for all words or for only function or content words.[2]

Sentiment Analysis and Semantic averaging

In the methods described above, words are present or absent and their occurrence at different frequencies can indicate the intensity of their contribution. But if words are associated with numeric values by, for example, rating (or 'norming') them along some dimension, we can quantify documents by computing the average across words and with each word making a contribution. *Word norms* often involve having humans rate thousands of words along dimensions such as *valence* (positive or negative), *concreteness* (how easy is the word to visualize in your mind's eye). Averaged over a number of individual raters, the word norms become useful proxies for evaluating documents along the same dimensions.

The most popular of these approaches is *sentiment analysis*. Sentiment analysis uses word valences to compute the average positivity or negativity of a document. This is often used to quantify user reviews to produce a measure of popular approval for specific products (e.g., sentiment analysis of Twitter posts about a new movie or a new Apple product), or to reveal the emotional state of the author (e.g., sentiment analysis of presidential inaugural addresses or of social media users who might have mental health risks).

This is the method behind the *Hedonometer*, which is an online tool that tracks the day-to-day average happiness of Twitter in association with recent events (https://hedonometer.org), as well as other media (Dodds et al. 2011; Dodds and Danforth 2010). Because sentiment can be computed over time, statistical models can be fit to the temporal patterns to categorize characteristic rising and falling patters of plots in books and movies (Reagan et al. 2016; Del Vecchio et al. 2021). We can use a similar method to compute an historical *national valence index* based on the books published in different languages, to evaluate how people felt at different points in history. Thomas T. Hills et al. (2019) used this approach

[2] Function words, or *stop words*, are typically closed-class words like pronouns and articles that are often unrelated to the topical content. Content words are the non-stop words. We discuss this further below.

to show that sentiment correlated with measures of life-satisfaction since the 1970s and with economic and health indicators since the 1800s.

By using different norms, one can evaluate different historical patterns. T. T. Hills and Adelman (2015) used concreteness norms to ask whether American English is becoming more concrete or more abstract over the last several hundred years. Over this same time period, the *Flynn effect* documents rising IQ scores across many populations including the United States. Could it be that American English reflects a change in, for example, symbolic or abstract reasoning? As it turns out, American English has seen a dramatic rise in concreteness over this time period, in books, newspapers, and even presidential speeches. This is particularly meaningful because concreteness is a property of words that are more easily remembered and more interesting. Using other norms and historical indicators, a number of different hypotheses were tested regarding this change in concreteness, with the most well supported hypotheses being that concrete language is more competitive in an attention economy (T. T. Hills, Adelman, and Noguchi 2017; Thomas T. Hills 2019).

Dealing with changes in word meaning over time

One caveat of working with word norms over historical time is that the meanings of words may change. For example, words like 'risk' and 'gay' have undergone dramatic changes in meaning, valence, and frequency over the last 100 years (Li et al. 2019; Li, Hills, and Hertwig 2020). This is true of many other words as well, which means that word norms collected recently may not generalize to the language produced several hundred years ago. Several methods have been developed to deal with this. One is to include robustness checks that limit the analysis only to those words that are most stable over the period of interest. Thomas T. Hills et al. (2019) included such a check using only the top 25% most stable words in the languages they investigated.

An alternative method is to compute historical norms. This approach uses the words that words co-occurred with in the past to evaluate changing patterns in sentiment and semantics (Recchia and Louwerse 2015; Bullinaria and Levy 2007). *The Macroscope* is an online dashboard that will do this for for individual words, providing additional information about historical change as well, showing changing patterns in meaning over time (Li et al. 2019, see http://macroscope.intelligence-media.com/). Snefjella, Généreux, and Kuperman (2019) provide historical norms for concreteness computed in a similar fashion for many thousands of words.

Available Word Norms

Word norms already exist along a number of dimensions and for many different languages. Here is a sample list of some of the many different norms available:

- Age of acquisition (Kuperman, Stadthagen-Gonzalez, and Brysbaert 2012; Scott et al. 2019)
- Word frequency (Brysbaert and New 2009)
- Reaction time (Keuleers et al. 2012)
- Risk ???

- Valence (Warriner, Kuperman, and Brysbaert 2013; Scott et al. 2019)
- Arousal (Warriner, Kuperman, and Brysbaert 2013; Scott et al. 2019)
- Dominance (Warriner, Kuperman, and Brysbaert 2013; Scott et al. 2019)
- Concreteness (Brysbaert, Warriner, and Kuperman 2014; Scott et al. 2019)
- Humor (Engelthaler and Hills 2018)
- Imageability (Cortese and Fugett 2004; Scott et al. 2019)
- Familiarity (Stadthagen-Gonzalez and Davis 2006; Scott et al. 2019)
- Word association (De Deyne et al. 2019)
- Taboo (Janschewitz 2008)
- Sensorimotor strength (Lynott et al. 2020)
- Semantic size (Scott et al. 2019)
- Gender association (Scott et al. 2019)
- Emotion (Troche, Crutch, and Reilly 2017)
- Polarity (Troche, Crutch, and Reilly 2017)
- Social interaction (Troche, Crutch, and Reilly 2017)
- Morality (Troche, Crutch, and Reilly 2017)
- Thought (Troche, Crutch, and Reilly 2017)
- Time (Troche, Crutch, and Reilly 2017)
- Space (Troche, Crutch, and Reilly 2017)
- Quantity (Troche, Crutch, and Reilly 2017)
- Visual Form (Troche, Crutch, and Reilly 2017)
- Auditory (Troche, Crutch, and Reilly 2017)
- Tactile (Troche, Crutch, and Reilly 2017)
- Smell/Taste (Troche, Crutch, and Reilly 2017)
- Color (Troche, Crutch, and Reilly 2017)
- Gender (Scott et al. 2019)

This list is not exhaustive, but it should be sufficiently stimulating for those interested in taking this approach. Links to the data for many of these can be found here (https://aginglexicon.github.io/menu/norms.html), in association with (Wulff et al. 2019). In cases where you cannot find a set of norms or they are not currently large enough for your purposes, you might consider collecting them yourself. New methods make this fairly easy to do—using crowd-sourcing and best-worst scaling— even on a limited budget (Hollis and Westbury 2018; Engelthaler and Hills 2018).

Sorting and Machine learning

How do we identify what documents are about or the persistence of topical ideas over time? For example, if we are interested in the kinds of things people talk about when they talk about immigrants, it would not be sufficient to simply compute the valence, or some other semantic or psycholinguistic property (e.g., reaction time) of the words. More appropriate would be to use an unsupervised approach to identify topics, and then run analyses on these topics. In topic modelling, instead of creating dictionaries by hand, topics are generated from the data in an emergent fashion rather than from the words that researchers believe theoretically represent that category, hence reducing another potential source of bias.

A popular method to achieve this is *latent Dirichlet allocation* (LDA), also known as *probabilistic topic models* (Griffiths and Steyvers 2004; Blei, Ng, and Jordan 2003). In a nutshell, LDA is a hierarchical Bayesian model that models documents as a probabilistic distribution of topics, with each topic representing a probabilistic distribution of words. Thus, LDA sorts words into topics and topics in documents. This is a *data generating process* that models the human process of writing a document as picking topics, and then picking words within topics. There are off-the-shelf tools for LDA within R and Python and, given a sufficient amount of data, they can rapidly identify word and topic distributions for sets of documents.

One question the researcher must answer is how many topics are there (i.e., the dimensionality). The dimensionality (k) is a free parameter and can be answered in several ways. LDA topics can be thought as the resolution of a microscope (Barron et al. 2018; Nguyen et al. 2020): if a fine-grained resolution is required, then a large number of topics is better; if the number of topics is small, these topics become more general (Allen and Murdock 2021). The right number of LDA topics is determined more by the question than the data itself (Nguyen et al. 2020; Allen and Murdock 2021).

There are two general approaches to using LDA. One uses LDA as a dimensionality reduction tool, which then allows one to identify similar documents and the persistence of topic ideas over time. These cases typically require a dimensionality sufficient to capture the important variation. In practice, this is often on the order of a hundred. For example, Murdock, Allen, and DeDeo (2017) used this approach to examine the books read by Charles Darwin over a span of 23 years, to examine his patterns of exploration and exploitation around specific topics. Using a dimensionality of $k = 80$, they computed the topic distribution for each of Darwin's book and then measured how these differed from one another over time using *Kullback-Leibler divergence* (KL), a cognitively valid measure of surprise based on information theory. KL provides a measure of how surprising a new distribution is given the expectation of a previous distribution. (Murdock, Allen, and DeDeo 2017) found that Darwin's reading choices first exhibited a pattern of exploitation—choosing books similar in topics to previous books—then followed by exploration, with the topics diverging from books read in the past.

Barron et al. (2018) used a similar approach with LDA to track innovation over 40,000 parliamentary speeches during the French Revolution. Using $k = 100$ and KL to track divergence, they found that patterns of language use in the speeches tended to be innovative, but where more innovative by the left (e.g., Robespierre) than by the right (Abbé Maury). More novel speeches also tended to be highly transient, showing a more rapid reducing in KL over time. Using these two measures, they were then able to identify speeches that were particularly resonant, which had a particularly large impact on the future. For some additional insight into this approach as well as more words of wisdom on LDA, (Allen and Murdock 2021) is particularly insightful.

A second approach to using LDA is to reduce the dimensionality to a set of topics that can be individually interpreted. For example, (Li and Hills 2021) used this approach combined with semantic averaging to inform theories of intergroup contact around immigrant groups. They identified $k = 15$ topics that were meaningfully related to "immigration" in a multi-year corpus from the *New York Times*. These included topics like Crime, Terrorism, Books, Religion, and Restaurants and food, and Museums. They then showed how these topics were related to the views of different immigrant groups. They achieved this by creating an “immigrant” corpus, which contained words related to immigration, and then subcorpora that were specific to 60 minority groups (e.g., Greek, Palestinian, Brazilian, etc). Using word norms, they found a strong correlation between document concreteness and valence (more concrete language was also more positive). Then they further explored how different topics informed immigrant valence, with topics related to food and art associated with particularly positive views.

Li, Hills, and Hertwig (2020) used a similar method to evaluate the history of the word *risk*. Computing LDA over a 200 year period from the Google Ngram corpus, they identified $k = 15$ topics which were then used to provide a topical history of risk. From the 1800s to the 1950s the word ‘risk’ was largely associated with topics related to war and battle. In the 1950s, there is a rise in topics associated with economic risk, followed by nuclear risk, and then a growing period of health risks up until the present. Using counting and averaging methods, they also find an increasing frequency of the use of the word risk with a corresponding fall in valence. Thus, *risk* has become more prominent and more negative, even as life’s risks has diminished (e.g., Pinker 2011).

(Priva and Austerweil 2015) analyzed the topical history of articles in the journal *Cognition*, with several values of k, choosing a common set of topics for analysis that passed a set of criteria.

Often it is the case that the particular value of k is not too important, as long as it is large enough to capture the diversity of topics but not so large that the topics become uninterpretable from a psychological perspective. Some authors provide complementary analyses on other values of k as supplemental data supporting the main analyses (Murdock, Allen, and DeDeo 2017; Miani, Hills, and Bangerter 2021). In all the cases described above, varying the value of k within a small range tends to produce a similar set of topics. The *LDAvis* package in R is useful for visualizing these topics in an effort to identify those most likely to be meaningful.

There are efforts to produce algorithms for identifying the optimal number of topics. The *ldatuning* package in R, for example, provides a function to estimate k based on four validated algorithms. This is however rarely useful in practice, because the algorithms are not designed for psychological interpretability. It nevertheless frees researchers from making arbitrary choices that may be difficult to replicate. This might be the case when the variable under examination is k (when comparing corpora). In such cases, the k value should be obtained in a consistent and systematic way across corpora.

General Work Flow and Strategies for NLP

Below we provide a short summary of thoughts one should consider when doing research in psychologically-informed NLP. Following that, we provide some additional information on text-preprocessing.

A checklist for NLP

1. Identify the question you are trying to answer. How does X affect Y? What are the differences between What theory informs it?
2. How will you measure this in language? Define your terms—operationalize them—to determine how you will detect your question's features in language. For example, if you are measuring changes in the meaning of 'love' over the past several hundred years, how will you measure 'love' in language?
3. What methods will answer your question? Decide whether you will use a dictionary method, an existing set of norms, attempt to identify topics, all of the above, or something else. Do you need to develop your own dicitonary or norms? Will you just count the word *love*, or words with similar semantic meaning, or compute a Love Index based on a set of 'love norms'? Will you compute a set of topics associated with love based on the words that co-occur within a certain window around love-related terms?
4. What corpora will address your question? Identify what corpora are most appropriate. Is the data available to collect your own corpora? Use multiple independent corpora where possible. The Touchstone Applied Science Associates corpus might be inappropriate for 'love', but a database of historical poetry might work well.
5. How will you pre-process the corpora? Typical decisions include whether or not to include function words (e.g., 'the', 'those', 'of'). Whether or not to stem words (should 'cat' and 'cats' be considered the same word? What about 'is', 'was', and 'were' and other forms of the 'to be'?). We discuss this in detail below.
6. Identify alternative hypotheses for the potential changes you may observe. Then identify additional methods to rule these out if they become relevant.
7. Where possible, pre-register your study in the Open Science Framework. This will help you think more clearly about the approach you intend to use.
8. Once you have results, share them with others and collect additional alternative hypotheses. Find ways to evaluate them. This well help establish the weight of the evidence for the different hypotheses and guide future research.
9. In your write-up, favor transparency. Detail your motivation, operationalization, corpora, pre-processing steps, and so on. Even if much of it goes in supplementary material, your readers and your future self will appreciate your thoroughness.

Pre-processing pipeline

Despite the imperative need of cleaning raw texts to build corpora, a quick survey of the published literature reveals that there are no standardized protocols for text pre-processing. A possible explanation for this lack is because pre-processing is deeply intertwined with the project's research question(s) and methodology, hence highly variable and idiosyncratic. Nonetheless, we sketch a non-exhaustive coarse-grained pipeline in the following steps: 1) word segmentation; 2) converting to lower case; 3) text cleaning by removing punctuation, numbers, and symbols; 4) mappings, such as expanding contractions or converting British to American (or vice versa) English; 5) removing stopwords; 6) word stemming or lemmatization; 7) removal of idiosyncratic terms such as extremely infrequent words or misspellings; 8) Part-of-Speech tagging. Depending on the purpose(s) of the project, the researcher must thoroughly plan *whether*, *when*, and *how* to proceed with each step. In the following, we provide a brief overview of the pre-processing pipeline highlighting common issues and possible solution.

Word segmentation

Often, in NLP, words are the units of analysis, hence they need to be extracted from text. Word segmentation (or tokenization) is usually performed by separating words by space, at least in English. Sometimes, however, the individual word meaning highly depends on the context: *cold war* is neither a cold thing nor a war, but a historical period. Similarly, *Bush City* in Kansas, *George Bush*, and *green bush flower* are different types of *bushes*. A common solution to this problem during tokenization is to create *n*-ngrams, which are concatenations of *n* words. In doing so, *george_bush* is captured as a different token (or term in the vocabulary) than *green_bush*. A drawback of this procedure is that it is computationally costly causing an exponential growth of the corpus' vocabulary. A potential solution is to perform the ngramming after the removal of stopwords and lemmatization (if any) and then pruning (i.e., removing rare ngrams) the vocabulary.

Tokenization also splits compounds, which are concatenation of two or more words functioning as a multiword expression. Often, compounds alternate orthographic spelling between spaced (*car wash*), hyphenated (*car-wash*), and concatenated (*carwash*) formats. Historical computational linguists have shown that compound lexicalization favors the direct transition from spaced to concatenated spelling (Kuperman and Bertram 2013). Splitting hyphenated —or separating spaced— compounds into distinct tokens might results in unbalanced results when algorithms rely on bag-of-words raw input texts: *car wash* is not a car, yet it sums up instances of *car*. Extracting grammatical functions of words (i.e., part-of-speech tagging) helps aggregating compounds.

Lower casing

Texts (or tokens) are often converted to lower case, so that instances of words are treated equally regardless of capitalization (e.g., *The* and *the*). In many NLP applications, this could be done either before or after tokenization. One problem with lowercasing occurs when proper names and acronyms match other common words (e.g., *Bush*, *Apple*, *US*). Some bag-of-words algorithms such as LDA that extract co-occurrences within documents would be in general capable of associating the *Bush* president to politically-focused topics (often co-occurring with *president* or *George*) and the *bush* tree to vegetation topics.

If this distinction is important for the purpose of the project, then the researcher might think of parsing the text prior to tokenization, attaching to the word an indication of whether it is a proper name or a noun. Note that in some projects, words should maintain their original forms. This is the case, for example, when the goal is to match with norms that are case-sensitive (e.g., of frequency, Brysbaert and New 2009).

Text cleaning

Depending on the purpose of the project, some text components are not useful/informative, despite their abundance. This is often the case of punctuation when bag of words are needed (clearly, unless punctuation is a variable of interest). R packages such as *quanteda* allow researchers to remove punctuation, symbols, and numbers while tokenizing. In some cases, however, punctuation could be important in defining *genres*. This is the case, for example, of exclamation and question marks in distinguishing documents endorsing conspiracy theories (Miani, Hills, and Bangerter 2021).

Mappings

Sometimes it is useful to normalize English words to either British or American English (e.g., *colour/color*, *analyse/analyze*). Unless relevant for the project's purpose, there is no reason to consider the two spellings as different words. The R package *uk2us* contains 1,720 spelling correspondences. Other projects (e.g., https://github.com/HoldOffHunger/convert-british-to-american-spellings) cover more than 20,000 mappings that can be easily wrapped into an R or Python function.

Especially in informal texts, words are often contracted (e.g., *you'd* to *you would* or *gimme* to *give me*). In NLP, this could be a problem because *you're* is considered a different term than *you* and *are*. If, for example, pronouns or auxiliary verbs are important, then contractions should be expanded. The R package *qdapDictionaries* provides 70 mappings for contractions (e.g., *when's* to *when is*). The Wikipedia page (https://en.wikipedia.org/wiki/Wikipedia:List_of_English_contractions) provides a table with almost 200 mappings that can be extracted and wrapped into an R or Python function.

Stemming or Lemmatizing

In some cases, it is useful to reduce the corpus' vocabulary by stemming and lemmatizing, which are techniques that reduce words' inflectional forms to a common base form. While stemming is based on heuristics that trim the ends of words removing derivational affixes, lemmatization aims to achieve the same goal via the use of a vocabulary and morphological analysis of words. For example, the stemming of the words *operate*, *operating*, *operates*, *operation*, *operative*, *operatives*, and *operational* returns one term: *oper*. Differently, lemmatization offers a more fine-grained distinction between words' roots hence returning a more diverse set of lemmas: *operate*, *operation*, *operative*, and *operational*. In both cases, the token pool has been reduced, yet to different extents. Depending on the granularity of the vocabulary required, the researcher might choose whether to stem or lemmatize, using both or none.

Stemming and lemmatization should be decided in advance according to the analyses planned. Some word norms have been collected for lemmas (e.g., Kuperman, Stadthagen-Gonzalez, and Brysbaert 2012; Warriner, Kuperman, and Brysbaert 2013) and others for inflected words (e.g., Brysbaert and New 2009). LIWC overcomes this problem by adding an asterisk to the root of the word so to match all possible inflectional forms (e.g., *satisf** matches *satisfaction*, *satisfy*, *satisfactorily*, etc). The 194 built-in dictionaries in Empath, however, contain terms in the lemmatized form, hence, prior to using these dictionaries, words in the corpus should be lemmatized accordingly. A suggestion, prior to applying stemming or lemmatization, is to closely look at the dictionary or word norms, checking the form of the terms and run simulation with sentences created *ad hoc* to test the tools. For example, how terms such as *dog*, *dogs*, *dog123* respond to the term *dog* in the dictionary/norm?

Stopwords

Stopwords are usually non-content uninformative words such as *the*, *of*, *and*, *a* that have low discrimination power (Lo, He, and Ounis 2005). Several lists have been compiled but generally, the agreement is that stopwords are among the most used words in language. For example, according to Fry (2000), the top most used 25 words account for about one-third of all printed material in English; the first 100 for about half; and the first 300 make up about 65% of all printed material in English (note that Fry's top words are included in the *qdapDictionaries* R package). However, language use is different across genres, and as a consequence also stopwords (see e.g., stopwords for technical language, Sarica and Luo 2021; or for Twitter Saif et al. 2014). Some times, words from stopwords lists can be dropped when such decision is theoretically grounded. For example, in a corpus of conspiracy theories (Miani, Hills, and Bangerter 2021), there is reason to keep some stopwords, specifically those associated with in/out-group language (e.g., pronouns) and those associated with refutation and questioning (e.g., negation and wh- questions: *not*, *why*, etc). Similarly, using Twitter for extracting emotional content, the first-person pronoun *I* is crucial for identifying tweets associated with expressions of personal emotions through self-reports (Fan et al. 2019).

Pruning

Pruning a corpus is a common practice to save computational costs by removing extremely low- and high-frequency uninformative words. Several methods have been developed. Yang and Pedersen (1997) evaluated five different pruning methods: document frequency, information gain, mutual information, χ^2-test, and term strength. They found that document frequency (the number of documents in which a term occurs), the computationally simplest procedure, showed similar results to other more costly operations with a vocabulary reduction up to 90% of unique terms excluding stopwords. Another way of pruning a corpus is selecting terms extracted from other vocabulary. For example, if the goal is to use word norms, then, the corpus could be trimmed keeping only terms used in the norms (simply because documents' norm mean is computed for only terms in the norm). The R package *quanteda* offers full control for trimming a corpus. It allows researchers to remove terms above/below a certain threshold for both terms and terms-in-documents based on absolute counts, proportion, rank, and quantiles. For example, it is possible to remove terms that sum to less than N instances in the whole corpus (e.g., non-frequent proper names, brands, or locations). If during corpus pre-processing, it was decided not to use a specific stopword list, then it is possible to remove from the corpus the top-N ranked terms (most likely *the*, *of*, *and*, *a*, etc). *Quanteda* also allows researchers to prune a corpus from a selected list of terms.

Extracting Part-of-Speech

The Part-of-speech (POS) tagging —or parsing— extracts from each token in the corpus the corresponding grammatical category (e.g., verb, noun, adjective, etc, Marcus, Santorini, and Marcinkiewicz 1993) and syntactical relationships between words within a sentence. Extracting POS tags from text is a computationally expensive yet powerful tool in text preprocessing. POS tagging requires as input the raw unprocessed text because otherwise (e.g., without punctuation or stopwords) the process, based on probabilities, is compromised. For bag-of-words algorithms, extracting POS tags is useful as it disambiguates words by attaching the POS to the lemma. This process reduces vocabulary by lemmatizing tokens yet it is less aggressive in information loss than lemmatizing because the same lemma is coded with its category (e.g., *play* as either noun or verb).

For example, the sentence "*Mary went to a play where her brother played the role of George Bush behind the bush*", after tagging, becomes [Mary_NNP, go_VBD, to_IN, a_DT, play_NN, where_WRB, her_PRP$, brother_NN, play_VBD, the_DT, role_NN, of_IN, George_NNP, Bush_NNP, behind_IN, the_DT, bush_NN], where the tags NNP (proper noun, singular) and NN (noun, singular) distinguish the two *bushes*, and NN and VBD (verb, past tense) disambiguate the word *play* as either noun or verb. Such a method has been also employed by Google Ngram to disambiguate tokens (see e.g., differences between *cooking_VERB* and *cooking_NOUN*). In a similar way, in *Bill Gates bought an Apple product*, both *Gates* and *Apple* are tagged as proper names (differently from *Throwing an apple at the gates*).

Note that parsing is computationally costly in terms of processing time and storage space. The output from a parsed text is generally a table (i.e., a data frame) where each row corresponds to a token and columns store information about the parsed text such as IDs for documents, sentences and tokens, the actual tokens and their lemmatized form, POS tags, syntactical relationships, and entity (based on Named-entity recognition, e.g., *Bill Gates* is a person and *Apple* is an organization). There are several packages that extract POS from texts such as the R packages *spacyr* (a wrapper for the Python *spaCy*), *coreNLP* (a wrapper for the Stanford CoreNLP Tools), *koRpus* (a wrapper for the TreeTagger), and *udpipe* (whose pipeline processing is based on CoNLL-U version 2.0).

Cleaning from machine language

The pipeline described above takes for granted that documents are "clean", meaning that text is human readable. However, there are cases in which this condition is not met. For example, today, it is getting popular to build corpora by mining webpages from websites (see e.g., Miani, Hills, and Bangerter 2021; Baroni et al. 2009). Usually, the output of such operation is a text document formatted using a machine language, meaning that the usable (i.e., human-readable) text is encoded within HTML markups. In R, there are some packages that brilliantly facilitate the remove of machine language such as the packages *RCurl*, *XML*, and *rvest*. Other packages (e.g., the Python *beautifulsoup* and *Goose*) not only extracts text from HTML files but also remove noise and irrelevant text such as navigation links, header, and footer sections.

Impact of pre-processing on corpus and vocabulary size

As an example, we take the novel *Wuthering Heights* by Emily Brontë (1847) and show how text pre-processing affects the number of words (tokens) and the size of vocabulary in the corpus. Without any cleaning, the novel is composed of 142,625 words with a vocabulary of 10,274 unique terms. Lowering the case reduces the vocabulary to 9,776 terms. Removing punctuation and symbols reduced the corpus to 116,509 terms (vocabulary size: 9,759). Removing stopwords reduces it to 54,186 tokens, with limited impact on vocabulary, which was reduced by just the size of the stopword list (vocabulary size: 9,591). Lemmatization reduces vocabulary size to 7,036, while stemming reduces it to 6,286 unique words. Stemming after lemmatization further decreased vocabulary size to 5,988 unique terms. Pruning the vocabulary by selecting words that appeared at least five times within the whole novel reduces vocabulary size to 2,103 terms (1,996 terms after removing stopwords), while corpus size was reduced to 49,086 terms (after removing stopwords). In terms of processing time and storage space, using *quanteda* to pre-process the novel (tokenization, stopwords removal, and lemmatization) took 2.55 seconds to generate a token object of 945.7 kB, while parsing, via *spacyr*, took 50.36 seconds to generate a spacyr object of 13.4 MB. These numbers have to be taken as an illustrative example of the effect of pre-processing on vocabulary/corpus size. More systematic analyses on a larger (than $N = 1$) sample would reveal, without much surprise, similar figures.

Limits to inference

There are numerous challenges to implementing NLP, and still others specific to historical analysis. The people who do this work are more likely to be aware of them than most, and the literature is an archive of methods for dealing with these challenges. For example, methods of collecting and publishing text data change over time. In discussions with researchers at the British Library, they pointed out that something as simple as a change in font in a particular section of a newspaper could easily lead to decades of text data being left out of a sample simply because it couldn't be digitized by available methods. Thus, researchers should look closely at how their corpora are constructed and their limitations.

Data quality can also change for other reasons. For example, the Google Ngram corpus changes the kinds of texts that are being digitized over time, and careful investigation reveals this (Pechenick, Danforth, and Dodds 2015). Because secondary data, like language corpora, are often collected for reasons unrelated to your hypothesis, they may also change for reasons unrelated to your hypothesis. For example, in the computation of the National Value Index, cultural trends led to a rise in realism, and rising democratic tendencies led to less idealistic and often more negative documents. This is perhaps best indicated by the rising German sentiment during the Second World War, when much writing was state sponsored and censored (Thomas T. Hills et al. 2019).

None of these changes prohibit the use of historical language corpora—whether they do or do not will depend on the question. How changes in the background method of data collection might influence your hypotheses are themselves additional hypotheses, and they can sometimes be fairly dealt with (Richey and Taylor 2020; Thomas T. Hills et al. 2019).

An additional problem is semantic drift and general language change over time. Languages do change over time, even very short periods of time, and historical approaches are wise to keep this in mind (e.g., T. T. Hills and Adelman 2015). As noted above, using contemporary word norms to evaluate what texts meant in the past is problematic. But calibrating norms to historical co-occurrences can help address this problem (Recchia and Louwerse 2015; Bullinaria and Levy 2007).

Finally, it's worth noting that the notion of statistical significance is often useless with large data corpora. When you have thousands of data points, it can be difficult to *avoid* finding a significant result. This is the nature of null hypothesis significance testing—with more data one can detect increasingly small effect sizes. One solution is to compute effect sizes, compare the relative sizes of results for a variety comparisons to determine what matters most, and formulate clear hypotheses that lend themselves to results that pass the inter-ocular trauma test—they hit you between the eyes.

Many of the papers cited here provide excellent examples of how to control for and resolve these and other issues. Anyone doing work in this area should take a close look at the methodological details of some of this work to develop their intuition for the many ways that one can improve their inferential reliability. There are also more general 'guides' aimed at the providing insight into the 'text as data' approach (Jackson et al. 2021; Grimmer, Roberts, and Stewart 2022). Often good results depend on good theory. Resting one's hypotheses on good scholarship and backing up one's results with good experimental tests—especially when other researchers have already done them (whether they agree with your or not)—will lead to the most convincing research.

Summary

The methods of counting, averaging, and sorting provided here are not meant to be exhaustive. There are many additional methods that could be bent to one's purposes (e.g., semantic space modeling, embeddings, latent semantic analysis, and so on). Also not exhaustive are the methods we provide for developing a work flow and pre-processing. New methods and implementations are being developed daily. Many of the articles cited here also report additional measures for squeezing more information out of these approaches. This is a testament to the richness of NLP and its potential for modification to address novel questions. Indeed, the most challenging part of natural language processing might be question development. A well formed question will likely lend itself to one or more of the methods described here. If it doesn't, it is worth considering how to develop an algorithm to address it—it will likely be of interest to others.

References

Allen, Colin, and Jaimie Murdock. 2021. "LDA Topic Modeling: Contexts for the History & Philosophy of Science." In *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*, edited by Grant Ramsey and Andreas De Block. Pittsburgh University Press; Pittsburgh.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora." *Language Resources and Evaluation* 43 (3): 209–26.

Barron, Alexander TJ, Jenny Huang, Rebecca L Spang, and Simon DeDeo. 2018. "Individuals, Institutions, and Innovation in the Debates of the French Revolution." *Proceedings of the National Academy of Sciences* 115 (18): 4607–12.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Bollen, Johan, Marijn Ten Thij, Fritz Breithaupt, Alexander TJ Barron, Lauren A Rutter, Lorenzo Lorenzo-Luaces, and Marten Scheffer. 2021. "Historical Language Records Reveal a Surge of Cognitive Distortions in Recent Decades." *Proceedings of the National Academy of Sciences* 118 (30).

Bourke, Joanna. 2015. *Fear: A Cultural History.* Virago.

Boyd, Ryan L, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. "The Development and Psychometric Properties of LIWC-22."

Boyd, Ryan L, Kate G Blackburn, and James W Pennebaker. 2020. "The Narrative Arc: Revealing Core Narrative Structures Through Text Analysis." *Science Advances* 6 (32): eaba2196.

Brysbaert, Marc, and Boris New. 2009. "Moving Beyond Kučera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English." *Behavior Research Methods* 41 (4): 977–90.

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2014. "Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas." *Behavior Research Methods* 46 (3): 904–11.

Bullinaria, John A, and Joseph P Levy. 2007. "Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study." *Behavior Research Methods* 39 (3): 510–26.

Cortese, Michael J, and April Fugett. 2004. "Imageability Ratings for 3,000 Monosyllabic Words." *Behavior Research Methods, Instruments, & Computers* 36 (3): 384–87.

Davies, Mark. 2009. "The 385+ Million Word Corpus of Contemporary American English (1990–2008+): Design, Architecture, and Linguistic Insights." *International Journal of Corpus Linguistics* 14 (2): 159–90.

De Deyne, Simon, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. "The 'Small World of Words' English Word Association Norms for over 12,000 Cue Words." *Behavior Research Methods* 51 (3): 987–1006.

Del Vecchio, Marco, Alexander Kharlamov, Glenn Parry, and Ganna Pogrebna. 2021. "Improving Productivity in Hollywood with Data Science: Using Emotional Arcs of Movies to Drive Product and Service Innovation in Entertainment Industries." *Journal of the Operational Research Society* 72 (5): 1110–37.

Dodds, Peter Sheridan, and Christopher M Danforth. 2010. "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents." *Journal of Happiness Studies* 11 (4): 441–56.

Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter." *PloS One* 6 (12): e26752.

Engelthaler, Tomas, and Thomas T Hills. 2018. “Humor Norms for 4,997 English Words.” *Behavior Research Methods* 50 (3): 1116–24.

Fan, Rui, Onur Varol, Ali Varamesh, Alexander Barron, Ingrid A van de Leemput, Marten Scheffer, and Johan Bollen. 2019. “The Minute-Scale Dynamics of Online Emotions Reveal the Effects of Affect Labeling.” *Nature Human Behaviour* 3 (1): 92–100.

Fast, Ethan, Binbin Chen, and Michael S Bernstein. 2016. “Empath: Understanding Topic Signals in Large-Scale Text.” In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–57.

Fast, Ethan, Binbin Chen, and Michael S. Bernstein. 2017. “Lexicons on Demand: Neural Word Embeddings for Large-Scale Text Analysis.” In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 4836–40. https://doi.org/10.24963/ijcai.2017/677.

Fraser, Kathleen C, Jed A Meltzer, and Frank Rudzicz. 2016. “Linguistic Features Identify Alzheimer’s Disease in Narrative Speech.” *Journal of Alzheimer’s Disease* 49 (2): 407–22.

Frimer, Jeremy A, Karl Aquino, Jochen E Gebauer, Luke Lei Zhu, and Harrison Oakes. 2015. “A Decline in Prosocial Language Helps Explain Public Disapproval of the US Congress.” *Proceedings of the National Academy of Sciences* 112 (21): 6591–94.

Fry, Edward Bernard. 2000. *1000 Instant Words: The Most Common Words for Teaching Reading, Writing and Spelling.* Teacher Created Resources.

Greenfield, Patricia M. 2013. “The Changing Psychology of Culture from 1800 Through 2000.” *Psychological Science* 24 (9): 1722–31.

Griffiths, Thomas L, and Mark Steyvers. 2004. “Finding Scientific Topics.” *Proceedings of the National Academy of Sciences* 101 (suppl 1): 5228–35.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences.* Princeton University Press.

Hills, T T, and James S Adelman. 2015. “Recent Evolution of Learnability in American English from 1800 to 2000.” *Cognition* 143: 87–92.

Hills, T T, James Adelman, and Takao Noguchi. 2017. “Attention Economies, Information Crowding, and Language Change.” In *Big Data in Cognitive Science*, edited by M. N. Jones, 270–93. Psychology Press.

Hills, Thomas T. 2019. “The Dark Side of Information Proliferation.” *Perspectives on Psychological Science* 14 (3): 323–30.

Hills, Thomas T, Eugenio Proto, Daniel Sgroi, and Chanuki Illushka Seresinhe. 2019. “Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books.” *Nature Human Behaviour* 3 (12): 1271–75.

Hollis, Geoff, and Chris Westbury. 2018. “When Is Best-Worst Best? A Comparison of Best-Worst Scaling, Numeric Estimation, and Rating Scales for Collection of Semantic Norms.” *Behavior Research Methods* 50 (1): 115–33.

Jackson, Joshua Conrad, Joseph Watts, Johann-Mattis List, Curtis Puryear, Ryan Drabble, and Kristen A Lindquist. 2021. “From Text to Thought: How Analyzing Language Can Advance Psychological Science.” *Perspectives on Psychological Science*, 17456916211004899.

Janschewitz, Kristin. 2008. “Taboo, Emotionally Valenced, and Emotionally Neutral Word Norms.” *Behavior Research Methods* 40 (4): 1065–74.

Keuleers, Emmanuel, Paula Lacey, Kathleen Rastle, and Marc Brysbaert. 2012. “The British Lexicon Project: Lexical Decision Data for 28,730 Monosyllabic and Disyllabic English

Words." *Behavior Research Methods* 44 (1): 287–304.

Kuperman, Victor, and Raymond Bertram. 2013. "Moving Spaces: Spelling Alternation in English Noun-Noun Compounds." *Language and Cognitive Processes* 28 (7): 939–66.

Kuperman, Victor, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. "Age-of-Acquisition Ratings for 30,000 English Words." *Behavior Research Methods* 44 (4): 978–90.

Kyle, Kristopher, Scott A Crossley, and You Jin Kim. 2015. "Native Language Identification and Writing Proficiency." *International Journal of Learner Corpus Research* 1 (2): 187–209.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. "Content Analysis of 150 Years of British Periodicals." *Proceedings of the National Academy of Sciences* 114 (4): E457–65.

Le, X, I Lancashire, G Hirst, and R Jokel. 2011. "Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists." *Literary and Linguistic Computing* 26 (4): 435–61.

Li, Ying, Tomas Engelthaler, Cynthia SQ Siew, and Thomas T Hills. 2019. "The Macroscope: A Tool for Examining the Historical Structure of Language." *Behavior Research Methods* 51 (4): 1864–77.

Li, Ying, and Thomas T Hills. 2021. "Language Patterns of Outgroup Prejudice." *Cognition* 215: 104813.

Li, Ying, Thomas Hills, and Ralph Hertwig. 2020. "A Brief History of Risk." *Cognition* 203: 104344.

Lo, Rachel Tsz-Wai, Ben He, and Iadh Ounis. 2005. "Automatically Building a Stopword List for an Information Retrieval System." In *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, 5:17–24.

Lynott, Dermot, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. "The Lancaster Sensorimotor Norms: Multidimensional Measures of Perceptual and Action Strength for 40,000 English Words." *Behavior Research Methods* 52 (3): 1271–91.

MacKrill, Kate, Connor Silvester, James W Pennebaker, and Keith J Petrie. 2021. "What Makes an Idea Worth Spreading? Language Markers of Popularity in TED Talks by Academics and Other Speakers." *Journal of the Association for Information Science and Technology* 72 (8): 1028–38.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics* 19 (2): 313–30. https://aclanthology.org/J93-2004.

Miani, Alessandro, Thomas Hills, and Adrian Bangerter. 2021. "LOCO: The 88-Million-Word Language of Conspiracy Corpus." *Behavior Research Methods*, 1–24.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." In *Advances in Neural Information Processing Systems*, edited by C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Müller, Karsten, and Carlo Schwarz. 2021. "Fanning the Flames of Hate: Social Media and Hate Crime." *Journal of the European Economic Association* 19 (4): 2131–67.

Murdock, Jaimie, Colin Allen, and Simon DeDeo. 2017. "Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks." *Cognition* 159: 117–26.

Nguyen, Dong, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. "How We Do Things with Words: Analyzing Text as Social and Cultural Data." *Frontiers in Artificial Intelligence*, 62.

Orimaye, Sylvester O, Jojo SM Wong, Karen J Golden, Chee P Wong, and Ireneous N Soyiri. 2017. "Predicting Probable Alzheimer's Disease Using Linguistic Deficits and Biomarkers." *BMC Bioinformatics* 18 (1): 1–13.

Pechenick, Eitan Adam, Christopher M Danforth, and Peter Sheridan Dodds. 2015. "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution." *PloS One* 10 (10): e0137041.

Pennebaker, JW, RL Boyd, RJ Booth, A Ashokkumar, and ME Francis. 2022. "Linguistic Inquiry and Word Count: LIWC-22." Pennebaker Conglomerates. https://www. liwc. app.

Pietraszkiewicz, Agnieszka, Magdalena Formanowicz, Marie Gustafsson Sendén, Ryan L Boyd, Sverker Sikström, and Sabine Sczesny. 2019. "The Big Two Dictionaries: Capturing Agency and Communion in Natural Language." *European Journal of Social Psychology* 49 (5): 871–87.

Pilgrim, Charlie, Weisi Guo, and Thomas T Hills. 2021. "Information Foraging in the Attention Economy." *arXiv Preprint arXiv:2107.12848*.

Pinker, Steven. 2011. *The Better Angels of Our Nature: The Decline of Violence in History and Its Causes*. Penguin.

Preis, Tobias, Helen Susannah Moat, H Eugene Stanley, and Steven R Bishop. 2012. "Quantifying the Advantage of Looking Forward." *Scientific Reports* 2 (1): 1–2.

Priva, Uriel Cohen, and Joseph L Austerweil. 2015. "Analyzing the History of Cognition Using Topic Models." *Cognition* 135: 4–9.

Reagan, Andrew J, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes." *EPJ Data Science* 5 (1): 1–12.

Recchia, Gabriel, and Max M Louwerse. 2015. "Reproducing Affective Norms with Lexical Co-Occurrence Statistics: Predicting Valence, Arousal, and Dominance." *Quarterly Journal of Experimental Psychology* 68 (8): 1584–98.

Richey, Sean, and J Benjamin Taylor. 2020. "Google Books Ngrams and Political Science: Two Validity Tests for a Novel Data Source." *PS: Political Science & Politics* 53 (1): 72–77.

Saif, Hassan, Miriam Fernandez, Yulan He, and Harith Alani. 2014. "On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter."

Sap, Maarten, Anna Jafarpour, Yejin Choi, Noah A Smith, James W Pennebaker, and Eric Horvitz. 2022. "Computational Lens on Cognition: Study of Autobiographical Versus Imagined Stories with Large-Scale Language Models." *arXiv Preprint arXiv:2201.02662*.

Sarica, Serhad, and Jianxi Luo. 2021. "Stopwords in Technical Language Processing." *Plos One* 16 (8): e0254937.

Scheffer, Marten, Ingrid van de Leemput, Els Weinans, and Johan Bollen. 2021. "The Rise and Fall of Rationality in Language." *Proceedings of the National Academy of Sciences*

118 (51).

Scott, Graham G, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno. 2019. "The Glasgow Norms: Ratings of 5,500 Words on Nine Scales." *Behavior Research Methods* 51 (3): 1258–70.

Snefjella, Bryor, Michel Généreux, and Victor Kuperman. 2019. "Historical Evolution of Concrete and Abstract Language Revisited." *Behavior Research Methods* 51 (4): 1693–1705.

Stadthagen-Gonzalez, Hans, and Colin J Davis. 2006. "The Bristol Norms for Age of Acquisition, Imageability, and Familiarity." *Behavior Research Methods* 38 (4): 598–605.

Troche, Joshua, Sebastian J Crutch, and Jamie Reilly. 2017. "Defining a Conceptual Topography of Word Concreteness: Clustering Properties of Emotion, Sensation, and Magnitude Among 750 English Words." *Frontiers in Psychology* 8: 1787.

Warriner, Amy Beth, Victor Kuperman, and Marc Brysbaert. 2013. "Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas." *Behavior Research Methods* 45 (4): 1191–1207.

Wulff, Dirk U, Simon De Deyne, Michael N Jones, Rui Mata, Aging Lexicon Consortium, et al. 2019. "New Perspectives on the Aging Lexicon." *Trends in Cognitive Sciences* 23 (8): 686–98.

Yang, Yiming, and Jan O Pedersen. 1997. "A Comparative Study on Feature Selection in Text Categorization." In *Icml*, 97:35. 412-420. Nashville, TN, USA.

Zenker, Fred, and Kristopher Kyle. 2021. "Investigating Minimum Text Lengths for Lexical Diversity Indices." *Assessing Writing* 47: 100505.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort.* Oxford, England: Addison-Wesley Press.