Computational History: Using Semantic Models to Measure Changes in Attitudes, Values, and Beliefs from Language

by

Matthew Cook

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

Department of Psychology

University of Manitoba

Winnipeg, Manitoba

Abstract

Language is not only a tool for communication, but a window into human nature and the mind. The way we talk about ourselves, others, and the world around us reveals our personalities, mental health, self- and group-serving biases, and more. Though many sources of language exist to understand human psychology, newspapers provide a unique opportunity for studying changes in attitudes, values, and beliefs. In this thesis, I analyzed 221 years (1789-2009) of American historical newspaper data from two historical newspaper corpora (Chronicling America and the Corpus of Historical America English [COHA]). The volume of data is more than could ever be read by any scholar or group of scholars. Therefore, I made use of a standard computational language model of distributed semantics called the Random Permutation (RP) model that "reads" through a corpus of text and generates a mathematical representation of word meaning (i.e., a vector representation). I used the RP model to generate a vector representation for each word written in each decade of the newspapers. The result of this procedure is a 3000-dimensional vector space where each word in each decade is represented as a point in a space that evolves through time. Similar words (e.g., *dog* and *canine*) occupy similar regions in the semantic vector space, whereas dissimilar words (e.g., *dog* and *toolbox*) occupy dissimilar regions in the semantic space. Having derived dozens of sets of vectors, I first conducted a series of four experiments using unambiguous ground truths to validate the semantic meaning embedding within the vector space. After validating the vectors, I used several methods, including machine learning methods, to measure long term changes in attitudes, values, and beliefs through an analysis of language. In addition to the computational work, I conducted an empirical experiment that demonstrated that the methods I used to measure meaning also predict peoples' behavioural bias in real-world consequential decisions (i.e., job hiring). The ultimate goal of the thesis is to

advance computational methods for accurately predicting people's emotions, thoughts, and behaviour from language.

Keywords: computational humanities, computational social sciences, distributed models of semantics, natural language processing, text classification

Acknowledgements

I would like to thank my advisor Dr. Randall Jamieson for his long-term mentorship, patience, help in preparing this thesis, and for being instrumental in influencing the way I think about the world. Thank you to my PhD committee, Dr. Andrea Bunt, Dr. Lorna Jakobson, and Dr. Johnson Li for serving on my committee and providing valuable feedback and support. Thank you to Dr. Harinder Aujla for providing me computing resources. I would also like to thank NSERC, as this research was supported, in part, by an NSERC PGS-D grant. Last, but not least, thank you to my family and my wife Amber, for everything.

Table of Contents

List of Figures

Chapter 1: Introduction

Everyday our brains perform an endless number of wonders. Our brains regulate our patterns of sleep, hunger, and thirst. Changes in air pressure detected by our ears are translated by the brain into meaningful sounds, and light falling onto our two-dimensional retinas are translated into three-dimensional representations of the world. We easily recognize the faces of our friends and foes, navigate the complex norms and rules of our social worlds, and store and retrieve memories that guide our emotions, thoughts, and behaviours. However, of all the tools in our cognitive toolbox, none is more powerful than language.

From birth to adulthood, we learn the meaning and proper use of tens of thousands of words (Pinker, 2013). Yet, memorization is only the beginning of our linguistic mastery. Our ability to understand and produce language allows us to combine words into endless novel utterances to communicate our thoughts with others (Pinker et al., 2003). Every society on Earth relies on language for the transmission of culture, education, science, and technology. Most impressively, language comes naturally to us. As Charles Darwin noted, “Man has an instinctive tendency to speak, as we see in the babble of our young children, while no child has an instinctive tendency to bake, brew, or write” (Darwin, 1902).

Language as a Window into the Human Mind

For psychologists, natural language offers a window into the workings of the human mind. Children’s mistakes in acquiring language, such as when a child says “she goed away”, reveals how our brains detect subtle patterns in our environment and apply abstracted rules (e.g., Marcus et al., 1992). The language we use in describing ourselves and others reveals our understanding of the social world and group dynamics (e.g., Gustafsson et al., 2014). We describe some unrelated strangers as our countrymen, our brothers and sisters. For other groups

of people who we feel less allegiance with, less polite terms are applied. As a final example, the way we talk about our emotions, thoughts, and behaviours serve as the basis by which mental health professionals assess our personalities and mental health (Balogh, et al., 2015 provides a broad overview of the importance of language in mental and physical health diagnosis, with Meehl, 1954 pioneering methods of automated computational and statistical diagnosis).

There are endless possible sources of language that psychologists can use to understand individuals or groups of people, and most of these come from written text: Personal journal or diary entries, published written works such as books, newspapers, or magazines, text messages, web searches, therapeutic or judicial transcripts, comments on social media, or a message scribbled on a sticky note. By analyzing a person's language, we can understand their background knowledge and assumptions, intuitions of cause and effect, judgements of positivity or negativity, and associations and concepts that remain hidden without observing their use of language (e.g., Khoo et al., 2002; Lindquist et al., 2015). These text sources allow psychologists to better understand cognitive and emotional development, mental health, personality, group dynamics, close relationships, and more. In essence, language can be used to understand every aspect of psychology. Indeed, these areas of inquiry form the basic sub-disciplines of psychology (cognitive, clinical, developmental, personality, social, etc; see Holtgraves, 2013 for a review within social psychology and Marcus et al., 1992 for an example from child development).

These same opportunities to understand humanity through our use of language apply to people who are no longer alive. Even though we cannot converse with people from the distant past, they have left a written record that we can excavate to uncover their understanding of the natural and social world, their associations, and conceptions. Furthermore, by conducting a

historical analysis of language throughout time, we can investigate and understand the social and ethical norms, culture, politics, associations, and concepts from the past.

As psychologists, we want a clear understanding of the past, not just in terms of a record of events, but an understanding of the way past people thought, felt, and related to the world around them. Were people of the past less or more racist compared to today? Did their attitudes towards men, women, children, homosexuals, minorities, animals, or criminals differ strongly from our current conceptions? In some sense, we know the answer to these questions already. But, by analyzing the language of past people, we can develop a fuller and deeper understanding of how people's attitudes, beliefs, values, and conceptions have changed throughout time to understand the social trajectories of which we are still a part.

Furthermore, by analyzing past use of language to understand how our attitudes, beliefs, values, and concepts has changed throughout time, we may be able to develop an accurate model of long-term psychological change. If we can accurately model these changes, we would not only have a fully descriptive account of how psychology has changed in time through language, but we could accurately predict future changes in society, culture, politics, and morality, even if now these ideas only seem like the stuff of science fiction.[1]

Text Analysis Methods

Many methods exist to analyze text (for reviews see Widdows, 2004; Jones et al., 2015; Turney & Patel, 2010). Traditionally, scholars have relied on close readings of historical texts. For relatively small bodies of text this works well. As an example, a scholar can set out to read all of Shakespeare's work in their career to uncover the themes, motifs, and choices of dialogue, setting, and plot used in the great master's work. However, especially for long-term historical

[1] In Isaac Asimov's Foundation series, the main character Hari Seldon develops a new mathematics for predicting human behaviour know as *Psychohistory*.

text analysis, this type of close reading is not a practical approach (Moretti, 2013). Often there is too much data for any single scholar, or even group of scholars, to read, interpret, and disseminate their findings to the larger research community.

Computational methods offer another approach to conducting large scale historic text analysis (Silge & Robinson, 2016). Rather than humans painstakingly reading through large bodies of text, a computer can be programmed to "read" through machine readable text (i.e., plain text). Many ways of analysing text have been developed, but the most basic text analysis method is frequency analysis. With frequency analysis, the frequency of a particular key word or phrase is documented throughout the history of a text corpus (typically a count of word occurrence relative to the total number of words in the corpus). For example, the frequency of the word "war" in a corpus of text would be expected to spike during periods of known warfare (e.g., the two world wars) relative to times of known peace. This frequency approach to text analysis is the basis for Google's text analysis interface Google Books Ngram viewer, which has been used for a great deal of text analysis research (e.g., Greenfield, 2013).

Though simple, frequency text analysis has several benefits. The method offers a fully automated method of text analysis for databases of any size, reducing the need for humans to conduct a close and time-consuming reading. The method offers a completely objective, repeatable, and formal method of analyzing any body of text. The method can be used to conduct within-group analysis, such as differences in word frequency in a single newspaper publication across time (e.g., The New York Times), or between-group analyses, such as the frequencies of particular words between two different newspapers (e.g., The New York Times compared to The Washington Post). The method also allows for mathematically representing documents of any size, whether the document represents a sentence or an entire book. By tallying the number of

times each word in a corpus appears for each document, documents are represented numerically (i.e., a vector), and documents can be compared for similarity using any standard indices of similarity (i.e., correlation, dot product, or cosine similarity).

Though there are an overwhelming number of papers that use the frequency approach, I will quickly point your attention to three papers that are especially related to this thesis as they focus on historical text analyses of social issues and concepts. These papers serve as my main motivational starting point for this thesis.

Michel et al. (2011) analyzed over 4 million digitized books which represents approximately 4 percent of all books ever published between the years 1800 and 2000. Their analyses produced some interesting results. Their analyses showed that the size of the English lexicon is growing and, how grammar evolves as certain irregular verbs like *found* supersede regular verb alternatives like *finded*. They also conducted a series of analyses demonstrating that throughout history, society tends to forget past events quicker with each passing year, as well as demonstrating the same pattern for how past scholars (e.g., scientists, mathematicians, and physicists) as well as non-academic famous people (e.g., actors, writers, politicians) are forgotten quicker with each passing year. Lastly, the researchers presented a series of analyses showing that during the 1930s and 40s, suppression of certain Jewish authors could be detected in German texts compared to English texts. In short, this paper made a strong case for using large-scale text analysis for analyzing culture, an approach the researchers coined *Culturomics*.

Landsdall-Welfare, et al. (2016) conducted a large-scale frequency-based text analysis of 150 years of British newspapers (from 1800 to 1950). Critically, the researchers demonstrated that their analysis of newspaper corpus was better able to detect many historic events than more traditional book-based corpora, such as the Google Books N-Gram corpus used in the landmark

Culturomics paper by Michel et al. (2011). Their analyses shows that events like major wars, coronations, elections of new popes, as well as major outbreaks of diseases like cholera, influenza, smallpox, and the plague can be identified by frequency-based approach. They conducted many other analyses, demonstrating the death and birth of certain words and phrases, like the term *British* overtaking the term *English*. They also presented a series of analyses on values and beliefs, technology and the economy, and social change and popular culture. Lastly, they conducted an analysis that, consistent with popular knowledge and past research, showed that men are mentioned about in the newspapers much more frequently than women, across every time period analyzed.

Following up on Landsdall-Welfare, et al. (2016) gender bias analysis, Johns and Dye (2019) conducted a large series of analyses investigating gender bias in text. The researchers used a large database of the most common male and female names published by the United States Social Security Association to measure the relative use of male versus female names across a wide range of analyses. By analyzing a large corpus of published books, their analyses showed that male names occurred a significantly greater number of times than female names. This result persisted regardless of whether the book was fiction or non-fiction, across different genres (e.g., romance, horror, crime, documentary), whether the author was male or female (though female authors were more egalitarian), and regardless of the year the author was born, or the author's country of origin.

These papers are important for several reasons. Each paper analyzed large corpora of text from the past 150-200 years, detecting measurable trends in what I believe are some of the most important areas of human life. Each of these papers used automated text analysis methods that are reproducible and formal, allowing researchers to ask and answer questions not possible by

traditional close readings of text by humans. These papers provide a further reminder that our use of language and the way it evolves is not random, but rather that language has a measurable structure, and that this structure reveals the structure of human thought. Analyzing language allows psychologists to understand attitudes, beliefs, values, and concepts, and how they change throughout time.

However, whereas the frequency-based approach offers a good starting point for understanding psychology through language, I think more advanced methods are needed for truly understanding psychology through text analysis. Many questions are left unanswered by these analyses. For example, men are discussed more in media, but how are men portrayed differently than women? Has our attitude towards war changed throughout time, and is there a way we can measure how justifiable or regrettable wars were viewed to be in the past?

I am going to retain much of the methods used in these papers while trying to push their analyses further. Though frequency analysis is a powerful method for conducting text analysis, there are several limitations. Most notably, the method doesn't capture language's most important, seemingly unquantifiable, quality – meaning. Though we may be able to determine the use of particular words or phrases with the method, words are treated in complete isolation with no regard to which words are related to other words or how patterns of frequency may change together. This is troubling for at least two reasons. First, words can be different, but mean the same or similar things (e.g., *dog* and *canine*). Second, the same words can be used with different meanings. To a psychologist, *depression* is a diagnosable disorder, to a geologist it is an alteration in the Earth's surface, and to an American historian it is a period of economic decline starting in the 1930s. However, the frequency approach treats the word *depression* as the same regardless of whether it appears in the context of clinical psychology, geology, or history.

Similarly, the frequency approach treats the word *depression* the same throughout time even though our concept of what *depression* is has changed drastically throughout time and is still currently evolving.

Additionally, the frequency-based method is especially troublesome when it comes to representing and comparing documents. Because documents are represented as a count of all words of interest, the size of the representation of documents grows with each new word in the corpus. With large corpora, the representation of documents can become unwieldly. For example, a corpus with 100,000 words needs 100,000 numbers to tally the counts of each word to represent a document. Furthermore, large vectors can cause problems related to collinearity with some statistical and machine learning models (e.g., regression), especially when the vectors contain highly correlated dimensions.

Improving the Frequency Based Approach

Some researchers have sought to correct these two main problems with the frequency approach (no representation of meaning, and large vector dimensionality). Pennebaker and colleagues (e.g., Tausczik & Pennebaker, 2010) have developed a computational text model called LIWC (Linguistic Inquiry of Word Counts). As the name suggests, LIWC is a word count (i.e., frequency-based) method of text analysis. Instead of representing a document as a count of tens or hundreds of thousands of unique words, LIWC has approximately 70 word categories such as personal pronouns (e.g., I, me), anger words (e.g., hate, angry), and religious words (e.g., god, pray). LIWC solves the problem of semantics by introducing these linguistic categories that capture meaning, while also reducing the dimensionality of the vectors. LIWC has been used in many text analysis and classification tasks, such as diagnosing mental health (Eichstaedt, 2018),

classifying personality (Yarkoni, 2010), and detecting honesty and deception (Newman et al., 2003).

Though LIWC is a popular and successful text analysis tool, I believe the method falls short in several respects. Though an improvement over word frequency-based analysis, LIWC's notion of semantics is too crude. Language is far too complex and subtle in its meanings to be conveyed in 70 (human crafted) categories. Because the method is a top-down deductive approach (humans decide which words belong in which categories and what those categories should be) rather than a bottom-up inductive approach (the data show us what words belong together and how the words cluster), the tool is too inflexible for many text analysis tasks that heavily depend on the meaning of language and not just which words are used. Thus, the next step requires a more sophisticated model for understanding meaning.

Formal Models of Semantics

The goal of developing a formal method of semantics started with Osgood in the 1950s. Osgood (1952) had participants rate words on scales with varying dimensions such as *good-evil* or *valuable-worthless* called the Semantic Differential. Each word was rated on 10 such scales, allowing each word to be represented with just 10 numbers. These word vectors allow for the formal mathematical comparison of words. Semantically similar words, like *knowledge* and *wisdom*, share similar ratings on the scales such as *good-evil* or *valuable-worthless*, whereas dissimilar words such as *knowledge* and *hatred* share fewer similarities. Each word Osgood had participants rate therefore occupied a 10-dimensional vector space (though it is helpful to just think in two dimensions, like a two-dimensional scatter plot). Words that contain similar ratings (and meanings) occupy similar regions of space, whereas less semantically similar words are

more distant to each other. The figure below shows a semantic space in just two dimensions for 9 words.

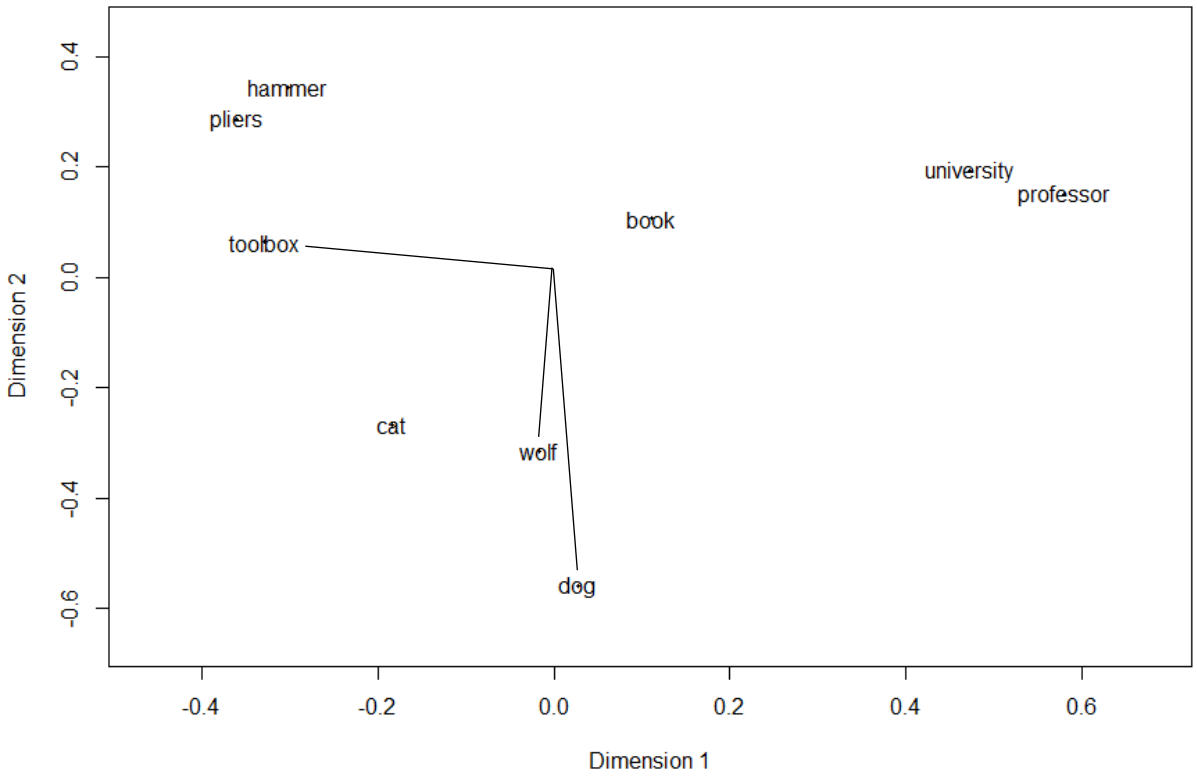


Figure 1.1 Multidimensional scaling solution of nine words plotted in two dimensions. In the space, semantically similar words (e.g., dog and wolf) are close together, whereas semantically dissimilar words (e.g., dog and toolbox) are farther apart. Alternatively, the angle between vectors departing from the origin (0, 0) can be used as a measure of their similarity.

However, like the close readings of historians, Osgood's Semantic Differential and the LIWC method of building meaning is time consuming, requiring many participants to derive stable estimates of ratings for each word of interest. Furthermore, as new words enter the lexicon (like *selfie*) or new uses of text emerge (like the use of emojis), new data need to be collected to

select or build categories for those new words. Starting in the 1980s and 1990s, psychologists began developing automated methods for building Osgood's semantic vectors (e.g., Lund & Burgess, 1996; Landuer & Dumais, 1997). Critically, these methods took a statistical or machine learning approach to deriving semantics where word meanings and relationships are learned from the structure of the data, rather than a priori notions of what constitutes word meanings. These methods quickly produce a mathematical model of word meaning. The most popular of these methods is Latent Semantic Analysis.

Computational Models of Semantics

Latent Semantic Analysis (LSA; Landuer & Dumais, 1997) is the first-generation machine learning semantic model. LSA learns word meanings from direct co-occurrence patterns in language. For example, the theory induces that *dog* and *cat* are semantically related because they appear together often and in similar contexts (i.e., sentences, paragraphs). The model also learns from indirect co-occurrence patterns. For example, *dog* and *canine* may rarely co-occur in the same sentence, but they co-occur in the context of similar words in different sentences (e.g., *cat*, *run*, *bark*).

Formally, LSA builds a numerical representation of semantics by re-representing a large body of text, such as a decade of newspaper data, as a matrix. The rows of the matrix represent each unique word encountered in the corpus. The columns of the matrix represent each document of the corpus, such as an article. The cells of the matrix represent the frequency that a particular word occurred in a particular document. After a frequency weighting technique is applied to the matrix, a technique borrowed from linear algebra called Singular Value Decomposition (SVD) is applied (Strang, 1998; Martin & Berry, 2011) that leverages the statistical regularities to generate a reduced dimensionality representation of the original word-by-document matrix. Each row in

the reduced matrix is called a semantic word vector. The columns in the reduced matrix are uncorrelated latent factors that maximize the variability of the original dataset. Figure 1 demonstrates the result of this process with a small, contrived example of just nine words (though the space in Figure 1 was produced using multidimensional scaling rather than SVD, the resulting process is essentially the same).

The result of this process is a set of vectors very similar to Osgood's manually derived vectors. In Osgood's case, the word vectors consisted of 10 dimensions that were manually derived and themselves semantically defined to represent meaning. In the case of LSA, the word vectors consist of several hundred dimensions of statistically derived, and not semantically defined, dimensions to represent meaning. In LSA, the latent dimensions are those which account for the most variance (i.e., information) across all the patterns of co-occurrence. These dimensions carry no inherent human meaning (though some researchers have sought to investigate these dimensions for their human meaning; Hollis & Westbury, 2016).

Formally, SVD is a decomposition of a matrix into the product of three more fundamental matrices (e.g., Strang, 1998):

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}},$$

where $\mathbf{A}$ is the original word-by-document matrix, $\mathbf{U}$ is a matrix of eigenvectors of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, $\mathbf{\Sigma}$ is a diagonal matrix with the square root of eigenvalues of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, and $\mathbf{V}$ is a matrix of the eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. Superscript-T indicates the matrix is transposed (columns in the matrix become rows and vice versa). The eigenvalues in $\mathbf{\Sigma}$ are ordered from largest to smallest and the eigenvectors are arranged to match the order of their corresponding eigenvalues. By tradition, the largest 300 eigenvalues and their corresponding eigenvectors are used to reconstruct a least-

squares best approximation of the original matrix (Landauer & Dumais, 1997). The reduced dimensionality vectors are derived by

$$\mathbf{X}_r = \mathbf{U}_r\mathbf{\Sigma}_r,$$

where *r* is the number of dimensions in the word vectors (typically 300 dimensions, at least by practical tradition; Landauer & Dumais, 1997).

The result of applying SVD is a high dimensional vector space of words. In this space, words that share similar meanings (e.g., *dog* and *cat*) move into and occupy overlapping regions of space. Words that are unrelated (e.g., *dog* and *toolbox*) move out of and are further apart in the space. See Figure 1 for a visual representation of a semantic space reduced using multidimensional scaling in two dimensions.

LSA is the first-generation semantic model due to the ease with which semantic vectors can be derived with any corpus of text. The model has proven its ability to track linguistic judgements of humans across a range of tasks (e.g., Foltz, et al., 1998; Pincombe, 2004). However, because word vectors are derived by performing a dimension reduction technique (i.e., SVD) on a matrix containing patterns of co-occurrence, the order in which words appear are not represented by LSA.

Modelling Word Order

Despite the success of LSA, the theory neglects word order (i.e., the *bag of words problem*). Second-generation semantics models were invented to solve the bag of words problem: most notably, the theory of Bound Encoding of the Aggregate Language Environment (BEAGLE; Jones & Mewhort, 2007).

In contrast to LSA, in BEAGLE the semantics of words are not computed in one operation (i.e., SVD) after the model has "read" all the text in the corpus. Rather, the semantics

of a word are slowly built up over time as the model "reads" through the text in the corpus. In BEAGLE, each word in a corpus is initially represented by an environment vector, **e**, with values drawn from a Gaussian distribution with the parameters $\mu = 0$ and $\sigma = 1/\sqrt{n}$, where n is the dimensionality of the vector (Jones & Mewhort, 2007). Like LSA, BEAGLE encodes direct co-occurrence information (i.e., *dog* and *cat* are similar because they appear in similar contexts) and indirect co-occurrence information (i.e., *dog* and *canine* are similar because they each appear with a common set of words). The model accomplishes this by summing neighboring word vectors into a target word's semantic representation. The context vector for a word is the sum of the environmental vectors for all other words that co-occur with it in a sentence. Formally a word's context vector is derived as,

$$\boldsymbol{c}_i = \sum_{j=1}^{j=w} \boldsymbol{e}_j$$

where $\boldsymbol{c}_i$ is the context vector for word i, $\boldsymbol{e}_j$ is the environmental vector for word j, w is the number of words in sentence j, and the environmental vector for the word being encoded is excluded from $\boldsymbol{c}_i$.

A word's semantic representation $\boldsymbol{m}_i$, is updated by including the new context information,

$$\boldsymbol{m}_i = \boldsymbol{m}_i + \boldsymbol{c}_i$$

where $\boldsymbol{m}_i$is the memory vector for word i and $\boldsymbol{c}_i$is the context vector for word i.

For example, after reading the sentence *the dog bit the mailman*, the memory vector for *dog*, m_{dog}, is updated as $m_{dog} = m_{dog} + e_{bit} + e_{mailman}$, the memory vector for *bit* is updated as $m_{bit} = m_{bit} + e_{dog} + e_{mailman}$, and the memory vector for *mailman* is updated as $m_{mailman} = m_{mailman} + e_{dog} + e_{bit}$. High frequency function words (i.e., *the*, *and*, *to*) are excluded when deriving a word's

context representation. Summing the environment vectors in this manner causes the memory vectors for all words in the same sentence to become increasingly similar to one another as the vectors develop. Perhaps less obvious, the method also encodes indirect associations between words. For example, even if *dog* and *canine* do not co-occur in the same sentence in the corpus, the vectors representing these words will be similar by virtue of having common words summed into their representations (e.g., *bark*).

Unlike LSA, BEAGLE also encodes word-order information using a mathematical operator called non-commutative circular convolution that associates two neighbouring words $\mathbf{w}_1$ and $\mathbf{w}_2$ by collapsing their outer-product matrix to form a new vector of equal dimensionality (Jones & Mewhort, 2007).

Formally, circular convolution encodes two vectors, **a** and **b**, into a third vector **z**,

$$\mathbf{z} = \sum_{j=0}^{n-1} \mathbf{a}_{j\ mod_n} \cdot \mathbf{b}_{(i-j)mod_n} \quad \{\text{for i} = 0 \ldots \text{n} - 1\}$$

where **z** is the convolution of vectors **a** and **b**, **a**, **b**, and **z** all have dimensionality of *n*, and the subscript mod_n refers to the modulo operator. Noncommutative circular convolution allows for a distributed (as opposed to local) fixed-dimensionality representation of word-order information that is neither commutative, nor associative, but distributes over addition and preserves similarity (Plate, 1995).

To encode the word-order information in a corpus, BEAGLE encodes information about a word's use, summing the convolutions of *n*-gram chunks of neighboring words. Formally, the order vector of word *i* is computed as,

$$\mathbf{o} = \sum_{j=1}^{p\lambda-(p^2-p)-1} bind_{ij}$$

where **o** is the order vector for word *i* in the corpus, *p* is the position of the word in the sentence, λ is a parameter that defines the maximum number of neighbors a word can be convolved with, and $bind_{ij}$ is the convolution of word *i* with word *j*. Traditionally, λ is set to 7 (i.e., consistent with Miller's, 1956, famous number 7 ± 2; see Jones & Mewhort, 2007, for a complete description of the method).

To illustrate, consider the sentence "the dog bit the mailman" which is encoded as,

$$bind_{dog,\,1} = e_a \circledast \Phi$$
$$bind_{dog,\,2} = \Phi \circledast e_{bit}$$

$$bind_{dog,\,3} = e_a \circledast \Phi \circledast e_{bit}$$
$$bind_{dog,\,4} = \Phi \circledast e_{bit} \circledast e_{the}$$

$$bind_{dog,\,5} = e_a \circledast \Phi \circledast e_{bit} \circledast e_{the}$$
$$bind_{dog,\,6} = \Phi \circledast e_{bit} \circledast e_{the} \circledast e_{mailman}$$

$$bind_{dog,\,7} = e_a \circledast \Phi \circledast e_{bit} \circledast e_{the} \circledast e_{mailman}$$

where $\circledast$ denotes circular convolution and Φ is a universal placeholder used in the computation of order information for every word in every position in every sentence (i.e., constructed as a random vector in the same way as the environment vectors), such that $m_{dog} = m_{dog} + o_{dog}$. The effect of this procedure is that words with similar functions in language (e.g., nouns versus verbs) become more alike, in much the same way that words with similar meaning become more alike.

Finally, the word-meaning and word-order vectors are summed into a composite vector that represents the sum of the word's meaning and order information,

$$\boldsymbol{m}_i = \boldsymbol{c}_i + \boldsymbol{o}_i$$

where $\boldsymbol{m}_i$ is the composite memory vector for word *i*, $\boldsymbol{c}_i$ is the context vector for word *i*, and $\boldsymbol{o}_i$ is the order vector for word *i*. Typically, $\boldsymbol{c}_i$ and $\boldsymbol{o}_i$ will be length normalized so that each

contributes equally to the memory representation regardless of the context or order vectors' geometric length.

Computational Improvements to BEAGLE

BEAGLE improves upon LSA by including a second set of vectors that build a representation of word order. However, circular convolution, the operator for building the order vectors, is costly from a computational point of view, making analysis of some text databases impossible due to the time it would take to build the vectors (e.g., Recchia et al., 2015).

Recently, researchers have begun to develop alternatives for building representation of word order rather than via the circular convolution operator (e.g., Sahlgren et al., 2008; Recchia et al., 2015). The most popular BEAGLE inspired method is known as the Random Permutation (RP) method (Sahlgren et al., 2008). Like BEAGLE, the RP method builds context vectors by vector addition, but uses a slightly different environmental vector (known as a spatter code vector). Also in contrast to BEAGLE, the RP method uses a permutation operation to encode order information.

The permutation operation operates by generating a new order vector for a word based on its position in a sentence. However, rather than generating and storing a new vector for each word in each position that it may occupy in a sentence, the method uses the same environmental vector, permuting it a certain number of times given a word's position in the sentence. The outcome of this procedure produces a similar order vector to BEAGLE. Critically, the RP method allows for building both context and order vectors using truly large corpora (e.g., all of Wikipedia; Recchia et al., 2015).

The RP method builds word vectors by reading a sentence at a time. Like BEAGLE, the RP method generates an unchanging environmental vector for each word. The RP method uses a

spatter code vector that is nearly all zeros, except for a few randomly placed +1s and -1s, which allows for the encoding of different words, similar to one-hot or dummy coding frequency coding schemes. To build a representation of meaning for the sentence *the dog bit the mailman*, the environmental vectors for all other words in the sentence get added to the target word. For example, the context vector for the word *dog* = *the* + *bit* + *the* + *mailman*. To build a representation of word order, a target words' neighboring words environmental vectors are permuted and added to the target words' order vector. For example, using a window size of two neighboring words, the order vector for the word $dog = the^{-1} + bit^{+1} + the^{+2}$. The superscripts represent the number of times the environmental vector for a word is permuted. In practice, rather than truly randomly permuting vector elements, the +1 subscript indicates that each element in a vector is shifted one element to the right (with the last element wrapping around to the first element location). The +2 subscript indicates shifting each element two positions to the right, and negative subscripts indicate shifting each element in the vector to the left. Permuting the vector in different directions allows for the unique encoding of neighboring words based on their position from the target word and whether the word precedes or follows the target word. This allows researchers to use the Random Permutation model to not only measure word similarity with the context vectors, but to measure which words are most likely to precede or follow a target word.

The RP method has several benefits over previously described models. Unlike LSA, the RP method builds word vectors by processing a sentence at a time. This has two major benefits over LSA. First, rather than needing to build one very large sparse word frequency matrix (which may be too large for the computer's memory) the vectors can be slowly built up as the model "reads" through the text corpus. Second, vectors can be updated with more data without the need

to start from scratch every time – one simply adds to the semantic vectors with new data. LSA requires the entire matrix to be updated and SVD recomputed to have an exact solution, unless one is willing to use approximate methods of updating the frequency matrix (Martin & Berry, 2011). Unlike BEAGLE, the RP method builds both context and order vectors using only vector addition without a need for circular convolution. This allows the RP method to be used with truly massive datasets that also escape the *bag of words* problem by generating a representation not only of semantics, but also order.

Research Goals

Distributional models like LSA, BEAGLE, and RP have been applied successfully to a broad range of tasks (Widdows, 2004; Turney & Pantel, 2010) including modelling children's semantic memory (Denhière et al., 2008), modelling basic memory processes (Howard, Addis, Jing, Kahana, 2007), predicting human performance in word recognition tasks (Buchanan et al., 2001), assessing personality (Kwantes et al., 2016), assessing reading skills (Magliano, & Millis, 2003), assessing and improving text comprehension (Millis et al., 2007), automatic essay grading (Landauer et al., 2003), building semantic search engines (Aujla et al., 2018), and assessing mental health from natural language (Willits et al., 2018; Johns et al., 2018; Bedi et al., 2015).

Historical text offers a unique opportunity to understand shifts and standards in the structure of language in culture, society, and morality through an analysis of printed language. Past researchers have shown that the frequency-based approach can be used to study changes in culture, with other researchers starting to use semantic models to measure change and differences in word and concept meaning. For this thesis, I will retain the methods and techniques used by researchers, while improving on their methods. I am focused on using computational models of semantics. In contrast to others, I am making use of the tools and techniques of cognitive

psychology (i.e., the Random Permutation model of semantics) rather than using other models grounded in more standard machine learning models (e.g., word2vec; Mikolov et al., 2013) or probability based models (e.g., GloVe; Pennington et al., 2014). By concentrating on using models from cognitive psychology, my hope is that I will have a psychologically inspired and informed method of text analysis.

Broadly speaking, the thesis is part of the domain of *corpus analysis* (e.g., Aujla, 2021; Johns & Jamieson, 2019) that demonstrates that temporal and spatial subsets of corpora can better explain peoples' language behaviour than one large general corpora. Rather than tuning semantic language models to a broad corpus that aims to model the average performance, the method focuses on curating corpora subsets that better explain performance and judgements for people with varying semantic representations (e.g., as discussed previously, the meaning of *depression* differs drastically whether you are a psychologist, geologist, or historian). Analogously, by building vectors from many temporal subsets of the larger corpora, I expect to find strong differences in semantic word meaning that allow us to track changes in human language and thought. The ultimate goal of this thesis is to better understand the psychological problems of tracking and predicting how concepts and the language that stands in place for concepts have changed and are changing throughout our collective, social history.

Chapter 2: Data Collection and Vector Derivation

Overview

For my thesis, I used two large corpora of historical text data to conduct a psychological analysis of attitudes, values, belief, and concepts with an eye towards cultural, political, and moral issues. The goal of the thesis was to develop a deeper understanding of past peoples' thoughts, emotions, and associations, by analyzing language. Furthermore, from a technical standpoint, the goal of this thesis was to advance text analysis methods from a psychological perspective.

Though many disciplines, such as statistics, machine learning, and computer science are interested in the problem of text analysis and natural language processing, I believe psychology has a unique role to play. Because psychologists study the processes by which people perceive, learn, think, decide, remember, and know, we approach problems with data, theories, and models already in hand for analyzing attitudes, beliefs, and values. These methods provide a psychologically-inspired and psychologically-informed perspective on building artificial intelligence.

The structure of this thesis follows three major sections: 1.) data collection and vector derivation (present chapter), 2.) validating the vectors (chapter three), and 3.) determining the structure of the vectors (chapters four and five). I will discuss each of these sections in detail below.

Data Collection

Two major sources of text for conducting large-scale historic text analysis are books and newspapers. For this project, I have chosen to mainly (though not exclusively) use newspapers for two main reasons. First, whereas books have a publication lag, sometimes taking years from

the time when they were written to the time they are published, newspapers are typically written hours before publication. Because I am interested in measuring changes across time, the nearly non-existent publication lag of newspapers will allow me the best chance of measuring changes in meaning accurately (e.g., Landsdall-Welfare et al., 2016). Second, the content of books can describe times and places far from the publication date and location, whereas newspapers are typically about local, current events. Science fiction, or historical fiction or historical non-fiction books could muddy the waters of my analyses, whereas the content of newspapers are typically about the current and local events.

For this project, I used the Chronicling America newspaper text database. Chronicling America is a partnership between the National Endowment for the Humanities and the Library of Congress and provides digitized historical newspapers from the late 1700s into the mid 1900s. The project aims to digitize American newspapers from 1690 to present day, but currently has digitized many American newspapers from 1789 to 1963 (175 years). Though I would have loved to extend my analysis from 1789 to present day using solely newspaper data, this was simply not possible. I was not able to find any text database that contained enough newspaper data that covered several hundred years from the late 1700s to present day. Furthermore, I was also not able to even find a newspaper text database that could fill the gap in data from the 1960s to present day.

The second source of data I used was the Corpus of Historical American English (COHA) which has newspaper data from the mid 1860s to 2009. Unfortunately, compared to the amount of data in Chronicling America, this corpus does not have enough newspaper data to build high resolution semantic vectors. However, the corpus contains newspaper, fiction, non-fiction, and magazine text data. I used all four of these sources from the COHA corpus to bring

my analysis up to present day. Though I would have liked to have solely newspaper data for this work, it was simply not possible. For each experiment, I used data from the Chronicling America database to build and analyze the vectors from 1780s to the 1960s, and I used the COHA database for the 1810s to the 2000s.

As of writing, the Chronicling America database is downloadable from 2245 .tar (tape archive) files, each of which contains thousands of digitized newspapers in text (.txt) format. In total, there are 14,910,627 text files (roughly 600 GB of .txt files). I wrote a short program in bash that downloaded each of the 2245 compressed .tar files, extracted their contents (the text files), and indexed the location of each text file along with storing metadata on what year the newspaper the text file represented was from.

The COHA database is downloadable from 20 zipped files. I wrote a short program in bash that downloaded, unzipped the files, and indexed the file location along with metadata on the data year and type (e.g., newspaper, fiction, non-fiction, and magazine).

Vector Derivation

For each decade from 1780s to the 2000s, I built both context and order vectors using the Random Permutation method previously described. I wrote a program in Java that "read' through a maximum of 250 million words for each decade from 1780s to the 2000s. Appendix A shows the word counts for each decade of the both the Chronicling America and COHA corpora. I used a standard word list of unique words in the TASA corpus that contained 92,393 common words. I decided on this list of words because the TASA corpus is a very common corpus used to build semantic vectors in psychology, allowing for direct comparison with the vectors I derived. I built one context vector and one order vector for each word encountered in each of the two corpora. Words that occurred in the corpus that were not in the list of words were excluded.

In line with past research (Recchia et al., 2015) and personal experience using these vectors, I choose a vector dimensionality of 3000, with six non-zero elements in each environment vector, containing three +1s and three -1s. The order vectors were constructed with a window of two words on either side of the target word.

The procedure produced 58 sets of vectors: 19 sets of context vectors and 19 sets of order vectors for the decades from the 1780s to the 2000s using the Chronicling America database and 20 sets of context vectors and 20 sets of order vector for the decades from the 1810s to the 2000s for the COHA database. In total, the program I wrote "read" over 5 billion words of text, and built tens of thousands of semantic vectors for each of the 23 decades I had data for.

Validating the Vectors

The ultimate goal of this project was to model long-term changes in values, attitudes, and beliefs through the use of language via text analysis. However, it would be scientifically and ethically suspect to make historical social claims without first validating the many tools and methods used and developed. The purpose of my first four sets of experiments were to validate the computational methods against unambiguous ground truths.

Chapter 3: Validating the Vectors

As a first evaluation of the vectors I have built, I wanted to make sure that the vectors I derived mapped onto our intuitive sense of meaning. For example, the model, like humans, should appreciate that *knowledge* is more similar to *wisdom* than to *democracy* or *broccoli*. One standard method of testing word vectors is with a synonym test. The reasoning goes that, if the model has built a representation of word meaning, it should be able to correctly select a synonym for a word amongst a list of other word options.

Experiment 1: TOEFL Synonym Test

The TOEFL (Test of English as a Foreign Language) is one of the most popular synonym tests and was used in Landauer and Dumais's (1997) landmark LSA paper, as well as being a standard part of United States college admissions. In the TOEFL test, a target word (e.g., *enormous*) is presented to the computer or human test taker along with four other words (e.g., *decidedly*, *uniquely*, *appropriately*, *tremendous*). The test-takers are asked to select the word that is most synonymous with the target word in a multiple-choice style test. After being presented 80 such questions, the percent of correct responses in calculated. The model's choice is determined by computing the similarity between the target word and the four options and selecting the word with the largest similarity. Landauer and Dumais's 1997 LSA model scored 64.4 percent correct, which is significantly better than a 25 percent chance model (a model that chooses randomly). The TOEFL test provides a standard method of evaluating that the word vectors I derived for each decade are embedding with our intuitive sense of meaning using a well-defined ground truth.

Method. I wrote a program in R that tested the model's performance on the TOEFL. For the first iteration of the experiment, I used the Chronicling America vectors that were derived

from the 1780s. The model was presented a target word vector, its synonym word vector, and three lure word vectors. The cosine similarity was computed between the target word vector and the four word multiple choice options (the synonym and three lures). The model's choice was determined by selecting the word with the highest cosine similarity to the target word. The model was given the 80 TOEFL questions and my program recorded the percent of correct responses.

The program iterated though this procedure for all sets of decade vectors from the 1780s to 1960s for the Chronicling America corpus and all sets of decade vectors from the 1810s to 2000s for the COHA corpus. To analyze the effect of the decade the vectors were derived from on performance, I conducted a simple linear regression with the model's percent correct on the TOEFL as a function of decade for each corpus individually. I evaluated the context vectors and the order vectors. The results for the order vectors for all experiments in this manuscript can be found in Appendix D.

Expected results. I expected three results. First, I expected the vectors built from text from all decades to perform above chance levels of 25 percent. If the model (or a person) guessed one of the four options at random, their expected performance would be 25 percent. Similarly, if the model (or a person) guessed the same multiple-choice option (e.g., the first options) every question, their expected performance would be 25 percent. Second, because the TOEFL is a modern test, I expected vectors derived from text from later decades (i.e., 1950+) to perform better than vectors derived from text from earlier decades, though I expected all vectors to perform appreciably better than chance. I expected there to be a statistically significant relationship between decade and performance as measured by proportion correct. Third, I expected the context vectors to perform best followed by the order vectors, given past research (Sahlgren et al., 2008) and my own experimental findings. This test allowed me to verify that the

model is capturing basic human meaning with a standard test that uses an established ground truth.

Results. Figure 3.1 shows the results from Experiment 1. The plot shows the percent of questions the model scored correctly as a function of the decade of the corpus from which the vectors were derived. Each point represents the model's performance for a given decade and the points are colored according to the corpus the vectors were derived from. The solid line shows the fitted least squares regression line.

There are several results to note. First, the vectors of every decade perform better than the 25 percent chance model. Second, there is a statistically significant relationship between the decade and the percent correct, with vectors from later decades performing better than earlier decades (R^2 = 0.46, p = 0.001 for the Chronicling America vectors and R^2 = 0.35. p = 0.006 for the COHA vectors). I suspect the vectors for later decades performed better because the TOEFL is a modern test with words that were common in that later part of the 20th century. Because these words were used more often in the later century, the model was able to build a more stable semantic representation of the word. Some of the differences in performance can also be attributed to corpus size. The last decade of the Chronicling America corpus (1960) has only four years of data compared to a full decade like previous decades (except the 1780s). The results with the order vectors perform nearly identically to the context vectors presented below. Therefore, I decided not to make the comparison of the context and order vector a central focus of the thesis. Replications of all experiments with the order vectors can be found in Appendix D.

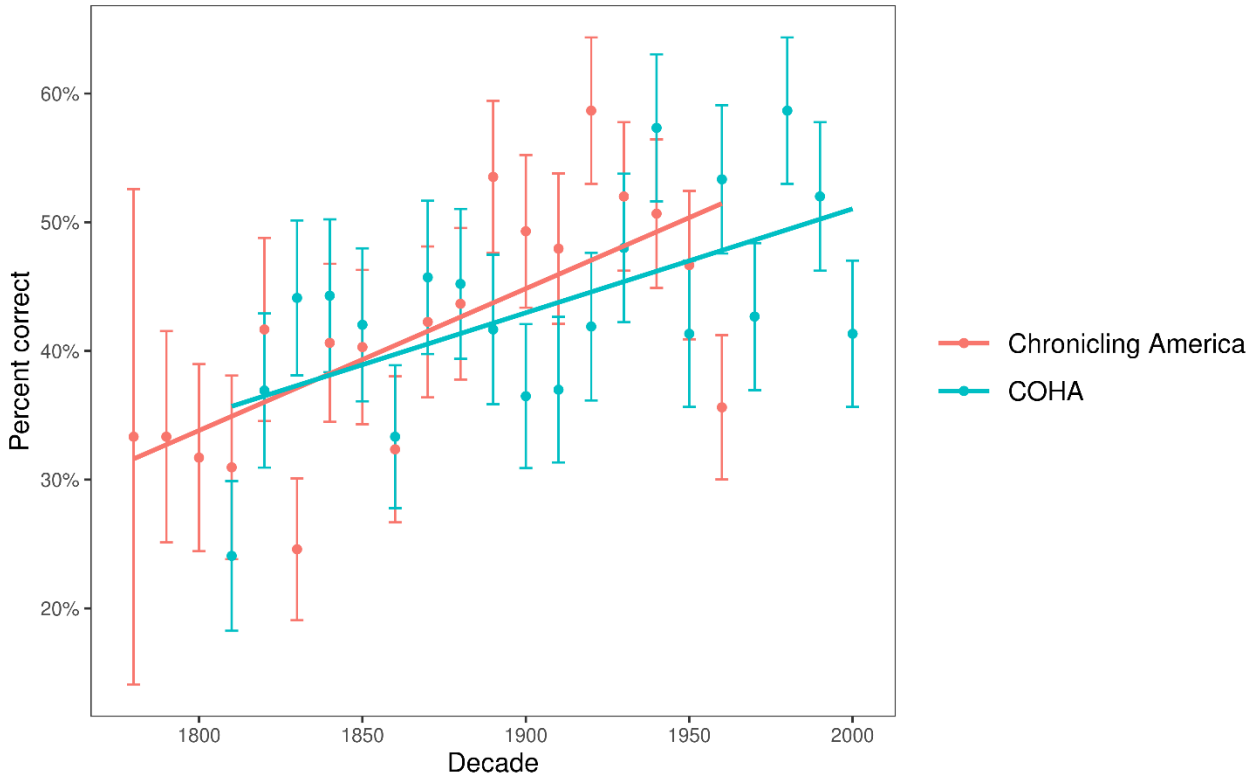


Figure 3.1 Results from Experiment 1. The x axis displays the decade and the y axis displays the percent correct. Each point represents the accuracy for a given decade on the TOEFL and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade. The error bars show the standard error of the proportion.

The results from this experiment show that the vectors I have derived carry meaning that allows them to perform well on a standard test of word similarity. Furthermore, the results show that the performance of the vectors on the modern TOEFL test increase when the vectors are derived from more current text that captures our intuitive sense of meaning. This suggests the possibility that language from the past differs in meaning from more recent texts given that the model struggled to perform on a modern test after being trained on data from the past.

However, there are many other ways to ensure the validity of the semantic word vectors I have derived. Another approach is to use word similarity tests.

Experiment 2: Word Similarity Tests

The TOEFL synonym test measures a model's ability to correctly identify a synonym of a target word amongst a group of four options. For thoroughness, I wanted to also test the model using another popular class of tests. Word similarity tests were devised to test a computational model's ability to match human word similarity data. In these tests, researchers first build an empirical database of human word pair rankings. Participants rate several hundred pairs of words (e.g., *dog-cat, dog-dictator, knowledge-dictator*) on a Likert-type scale and researchers compute the mean rating for each word pair. As one would guess, the *dog-cat* word pair would have an average higher similarity ranking from participants than the *knowledge-dictator* word pair. The result of this procedure is a database with average word pair rankings that describes the average semantic judgements of human participants. I wanted to ensure that the semantic vectors I derived also mapped onto this carefully established measurement of human semantic judgements for word pairs.

To test the computational model, the model is given the same word pairs as human participants provided ratings for and the model computes the similarity for each pair of word vectors. A measurement of similarity, such as a correlation or cosine similarity is computed between human ratings and the model's similarity values. I continued to measure similarity using the cosine similarity function. A high agreement (as measure by the cosine) between the model's similarity judgements and human similarity judgements indicates that the model's sense of word similarity is related to human's sense of word similarity.

Many word similarity tests exist, and it has become standard practice to test models using a good number of such tests (e.g., Johns et al., 2019). For rigour, I tested the model using six different word similarity tests: 1) MEN (*n* = 3000, Bruni, et al., 2012), 2) Mturk-771 (*n* = 771; Halawi, et al., 2012), 3) RG1965 (*n* = 65; Rubeinstein & Goodenough, 1965), 4) Simlex-999 (*n* = 999; Hill, et al., 2016), 5) Verb 143 (*n = 143;* Baker et al., 2014), and 6) WordSim *(n* = 353; Agirre, et al., 2009). The sample sizes reported for each test are the number of word pairs contained in each dataset. These word similarity tests provide a standard set of methods of evaluating that the word vectors for each decade are embedding with our intuitive sense of meaning using a well-defined ground truth.

Method. For the first iteration of the experiment, I used the Chronicling America vectors that were derived from the 1780s and used the MEN similarity test. I wrote a program in R that presented the model with each of the 3000 word pairs from the MEN word similarity database (e.g., *sun-sunlight, automobile-car, bright-grey*). The model then computed the cosine similarity between each of the 3000 pairs of words resulting in 3000 cosine similarity measurements. The program then computed the cosine similarity between the 3000 model similarity scores and the 3000 average human ratings of the same word pairs. The result of this procedure is one measurement of the average similarity between the model word pair judgements and human word pair judgements using the Chronicling America Corpus for the 1780s.

The program iterated though this procedure for all sets of decade vectors from the 1780s to 1960s for the Chronicling America corpus and all sets of decade vectors from the 1810s to 2000s for the COHA corpus using all six sets of word similarity tests. To analyze the effect of the decade the vectors were derived from on performance, I conducted a simple linear regression with the cosine similarity between the model and human similarity judgements as a function of

decade for each corpus individually. I evaluated the context vectors and the order vectors. The results for the order vectors for all experiments in this manuscript can be found in Appendix D.

Expected results. I expected a similar pattern of results for the word similarity tests as I expected for the TOEFL test. First, I expected the model's similarity ratings to produce a strong cosine similarity with human's similarity values across all decades. Second, though I expected high cosine similarity values for all decades, I expected performance with the vectors derived from more recent text (i.e., 1900+) to perform better than vectors derived from text from earlier decades. I expected there to be a statistically significant relationship between decade and performance as measured by cosine similarity. Third, I expected the context vectors to perform best, followed by the order vectors. Finally, with the RG1965 (Rubeinstein & Goodenough, 1965) I expected the vectors derived from the most recent text (i.e., from the 1960s) to perform best as these word similarity ratings are from the exact same time period as the text the vectors are derived from.

As with the TOEFL test, these tests allowed me to verify that the model is capturing aspects of word meaning with a standard test with a verifiable ground truth.

Results. Figure 3.2 shows the results from Experiment 2. The plot shows the cosine between the model's judgements of similarity and that of human judgements as a function of the decade of the corpus the vectors were derived from. Each point represents the model's performance for a given decade and the points are colored according to the corpus they were derived from. The solid line shows the least squares regression line.

There are several results to note. First, the majority of the vectors across decade perform significantly better than that would be expected by chance. Second, there is a statistically significant relationship between the decade and the model's cosine similarity with human

judgements, with vectors from later decades performing higher than earlier decades for several of the word similarity tests. Third, as hypothesized the RG1965 (Rubeinstein & Goodenough, 1965) word similarity test performed the worst.

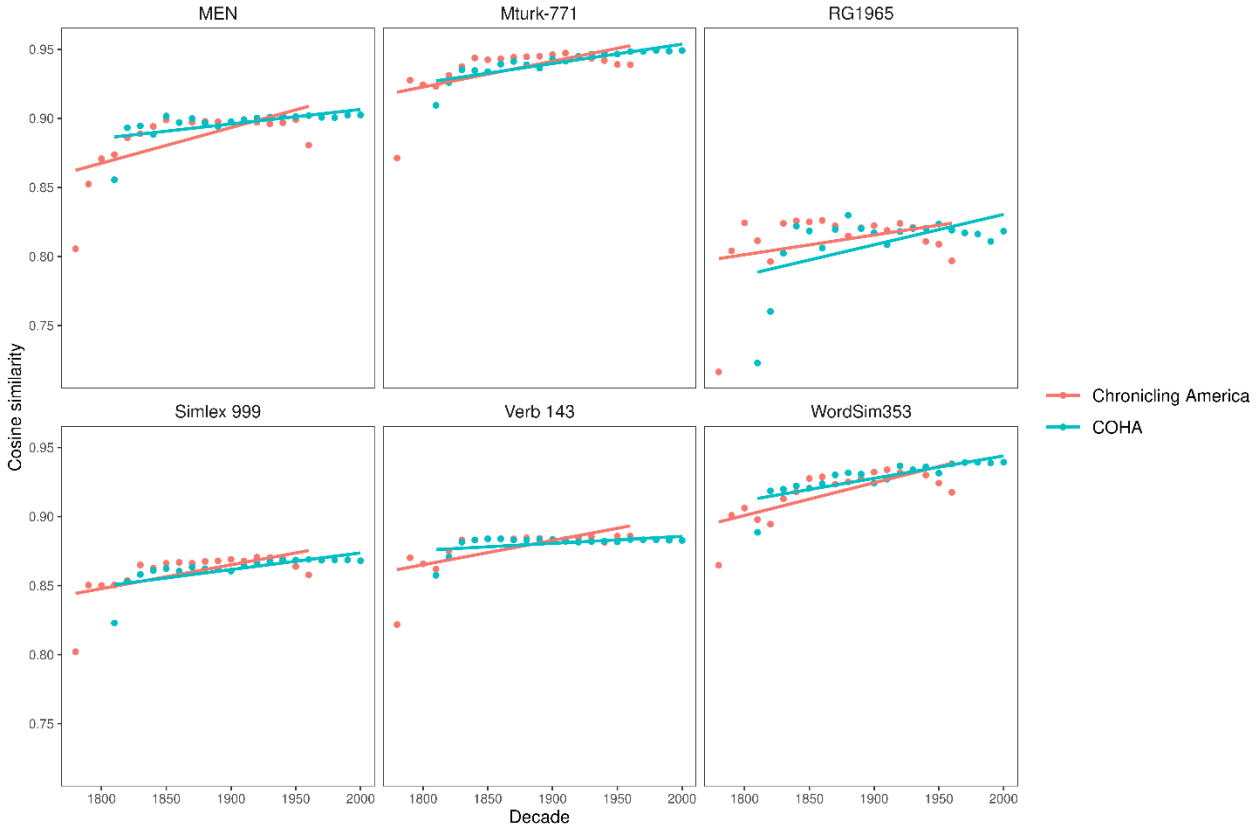


Figure 3.2 Results from experiment 2. In each of the six subplots, the x axis displays the decade and the y axis displays the cosine similarity for six different word similarity test. Each point represents the cosine similarity for a given test for a given decade and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade.

The results from this experiment and the previous experiment with the TOEFL show that the vectors I have derived carry meaning that allows them to perform well on two standard tests of word similarity. Furthermore, the results from both experiments show that the performance of

the vectors on the modern tests increase when the vectors are derived from more current text that captures our intuitive sense of meaning. I used one more standard method of validating the meaning of word vectors with a verifiable and unambiguous ground truth.

Experiment 3: Word Classification

Classification is a fundamental and standard part of natural language processing (Jurafsky & Martin, 2009), and is a large part of this thesis in the experiments that follow. To further ensure that the vectors I have derived map to human meaning and could be used in more complex classification tasks that follow, I tested the word vectors in a well-controlled word classification task.

For the word classification task, I built a set of 10 categories of words (*countries*, *animals*, *body parts*, *colors*, *sports*, *vegetables*, *fruits*, *diseases*, *professions*, and *furniture*) each containing 10 words. The complete word lists for this experiment can be found in Appendix B. The vector for each of these words is represented by one of 100 points in a 3000-dimensional space (though it is convenient to just think of a two-dimensional scatter plot with 100 points). Each point belongs to one of 10 categories. Ideally, these categories cluster together (i.e., all the animal words are close together in the space and distinct from the other 9 clusters). However, these categories can form more complex relationships. I employed one standard and popular classification model throughout the entire thesis. Random Forest models (Breiman, 2001) use a tree-based machine learning model called a decision tree. The model relies on information theory (Shannon & Weaver, 1949) to find a decision criterion to make its classification decisions. As a practical and intuitive example, in the domain of mental health, psychologists define a decision boundary and classify all patients as depressed when their score exceeds a threshold. Decision trees have the advantage that decision boundaries are derived automatically and empirically

using information theory to determine the optimal threshold. Furthermore, Random Forest models utilize a *wisdom of the crowds* approach, aggregating the decision of many decision trees each trained on different subsets of the data. Random forests are a standard classification model and are one of the most consistently best performing models across a range of classification tasks (Fernández-Delgado et al., 2014).

I tested the models' ability to classify novel words (i.e., words not part of the training process) using a standard Cross Validation method (Howell, 2010). To test the models' ability to correctly classify the word vectors, I used the standard *k* fold cross validation method that randomly splits the data into $k = 10$ even subsets. For each iteration of training, the task of the model is to fit the model on all subsets of the data except the k^{th} subset of data. To test the model, the model makes a classification decision of each example in the k^{th} subset. This procedure is replicated *k* times, with each of the *k* subsets being evaluated once. Importantly, each example in the dataset is only evaluated once which leads to an unbiased estimate of the model's ability to classify new examples. The goal of the model is to classify each example to its respective category without knowledge of its true category membership. I conducted this analysis for each decade from 1780s to 2000s using both the Chronicling America and COHA corpora.

This word classification task test provides a standard method of evaluating that the word vectors for each decade are embedding with our intuitive sense of meaning using a well-defined ground truth.

Method. For the first iteration of the experiment, I used the Chronicling America vectors that were derived from the 1780s. I wrote a program in R that split the 100 word dataset into 10 even groups (or folds to use the lingo of *k* fold cross validation). During each of $k = 10$ iterations, 90 labelled examples (i.e., word vectors with their associated category) are used to fit the

Random Forest model. The model is then evaluated by being presented the remaining 10 unlabelled examples (i.e., word vectors without their associated category) and the model makes it classification decision for each of the 10 unlabelled word vectors. The percent correct score is recorded. The result of this procedure is 10 accuracy scores that are averaged to produce an unbiased estimate of the model's ability to classify novel words to one of the 10 predefined word categories.

The program iterated though this procedure for all sets of decade vectors from the 1780s to 1960s for the Chronicling America corpus and all sets of decade vectors from the 1810s to 2000s for the COHA corpus. To analyze the effect of the decade the vectors were derived from on performance, I conducted a simple linear regression with the model's classification accuracy as a function of decade for each corpus individually. I evaluated the context vectors and the order vectors. The results for the order vectors for all experiments in this manuscript can be found in Appendix D.

Expected Results. I expected two results. First, I expected the vectors derived from text from all decades to perform above chance levels of 10 percent. If the model (or a person) guessed one of the ten categories at random, their expected performance would be 10 percent since there are 10 even sized categories. Second, I expected vectors derived from text from later decades (i.e., 1950+) to perform better than vectors derived from text from earlier decades, though I expected all vectors to perform statistically better than chance. I expected there to be a statistically significant relationship between decade and performance as measured by proportion correct.

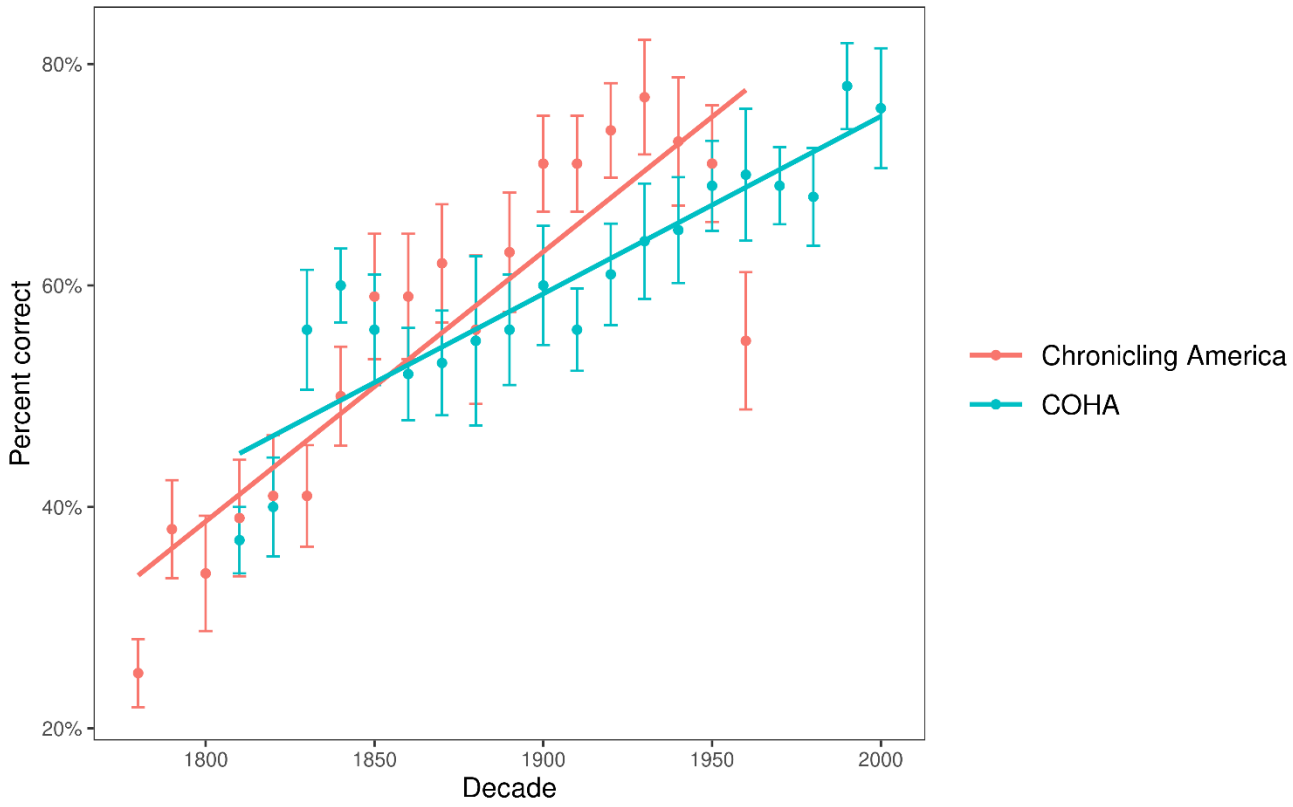


Figure 3.3 Results from experiment 3. The x axis displays the decade from which the vectors were derived and the y axis displays the percent correct for the word classification task. Each point represents the accuracy for a given decade and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade. The error bars show standard error of the mean.

Results. Figure 3.3 shows the accuracy in the word classification task as a function of the decade of the corpus the vectors were derived from. Each point represents the model's performance for a given decade and the points are colored according to the corpus they were derived from. The solid line shows the least squares regression line.

There are several results to note. First, all the vectors across decade perform significantly better than that would be expected by chance (i.e., 10% accuracy). Second, there is a statistically significant relationship between the decade and the model's correlation with human judgements, with vectors from later decades performing better than earlier decades ($R^2 = 0.78$. $p < 0.001$ for the Chronicling America vectors and $R^2 = 0.82$. $p < 0.001$ for the COHA vectors). Third, the vectors from both corpora perform approximately equivalent, though the Chronicling America vectors perform better during the early 1900s. Lastly, the vectors for the first and last decades (the 1780s and 1960s) of the Chronicling America corpus perform appreciably worse than the neighboring decades. This is likely due to the fact that there are not a full decade of data in the 1780s corpus (there is only data for one year – 1789) and not a full decade of data in the 1960s corpus (only data from 1960 to 1963).

Interim Summary

I have demonstrated that the vectors I have derived perform well on three different classes of tests: word synonym tests, six different word similarity tests, and a word classification task. These results show that the vectors I built from this historic database of text track human sense of word meaning. I consider these three tasks critical tests for the model to perform well before continuing with more complex social analyses. If the model had failed these basic tests, I would have no confidence in any further results going forward. I would therefore find it necessary to re-derive the vectors using more data, or a different set of hyper-parameters using the Random Permutation model (e.g., vector dimensionality, window-size, or the composition of the environment vectors). However, given the model performed well across these basic tasks, I went on to apply similar methods to investigate meaning in the social realm.

Chapter 4: Applying the Vectors to the Social Psychology Realm

Experiment 4: Detecting Hate Speech in Social Media

Testing the vectors using the TOEFL, six different word similarity tests, as well as a word classification task, has shown the model's ability to track human semantic judgements. However, these tests don't necessarily track more interesting differences in attitudes, beliefs, and values. Ideally, the vectors should be able to track social topics, such as the timely and relevant issue of hate speech. Though many definitions exist, hate speech is broadly defined as speech towards a minority or disadvantaged group that is intended to promote harm (Jacobs & Potter, 2000).

Many researchers have used computational models to identify hate speech (Kwok & Wang, 2013; Burnap & Williams, 2015; Djuric et al. 2015). Davidson et al. (2017) compiled a large database of almost 25,000 tweets that contained words typically associated with hate speech. However, not all tweets that contain such words are hate speech. For example, though using the word *bitch* may be associated with hate speech, people may simply be quoting a song lyric or some other context may make the use of the word justifiable and not hateful. The researchers had participants categorize each tweet as either *hate speech*, *offensive but not hate speech*, or *neither hate speech nor offensive*. They represented the tweets as vectors and classified the tweets using logistic regression. Their model performed very well (i.e., approximately 90 percent accuracy).

As another test of the vectors, I used Davidson et al.'s (2017) database of almost 25,000 tweets. I only used the tweets labelled *hate speech* and *neither offensive nor hate speech,* using equal number of both categories to establish a 50 percent chance model in a binary classification task. Informally, I built a representation of each tweet that captures a tweet's meaning. I then

used the same standard classification model (Random Forest) as the previous experiment to train the models to attempt to predict novel tweets. I wrote a program in R that recorded the percent correct that the model scores (i.e., what percent of novel tweets the model correctly identified correctly as either *hate speech* or *neither offensive nor hate speech*). Importantly, I evaluated the models on their ability to classify novel tweets, not to simply fit known data. I conducted this series of steps for each set of decade vectors I build from the 1770s to the 2000s.

Formally, I built a representation of each tweet by taking the vector for each word in the tweet and computed the average of those vectors. This new vector represented the average meaning contained in the tweet, even if that vector did not correspond to any single word in the tweet or corpus more broadly. The result of this procedure was a 3000-dimensional vector space of tweets. Ideally, similar tweets occupied close regions in the semantic space, and less similar tweets occupied more distant regions of space. Each point in the space belonged to one of two categories (i.e., *hate speech* or *not hate speech*). Ideally, these two categories were entirely distinct non-overlapping groups in the space. In practice however, the groups tended to form a more complex non-linear relationship (see the figure below for a visual demonstration of a simulated linearly separable classification problem and a simulated non-linearly separable classification problem in two dimensions). To learn the relationship between the location of the tweets in space and their associated categories, I used the same tree-based classification models as the pervious experiment to classify the novel data. Importantly, and because of the large amount of data I had available for this experiment, I randomly partitioned the data into a training set and a test set. I used the training set to estimate the model's classification accuracy using the *k* fold cross-validation method described previously. As a final test, I evaluated the model on the

test set, which was not used to inform any other training process. This approach is the gold standard method of evaluating models (e.g., Kuhn & Johnson, 2016).

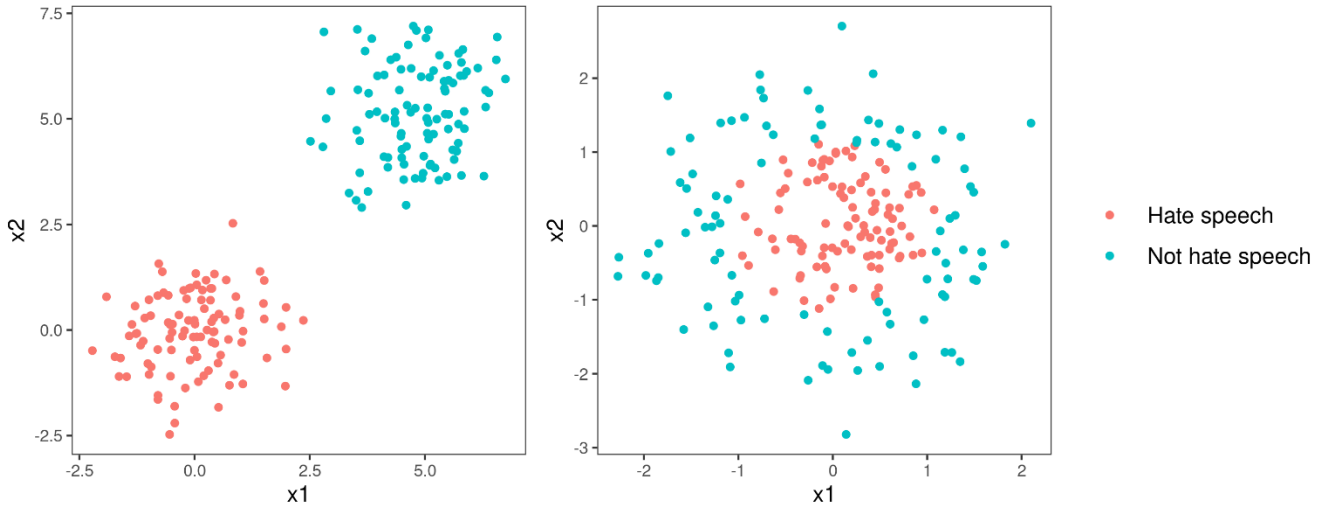


Figure 4.1 Simulated and simplified examples of a two-dimensional vector space of hypothetical tweets. The red points represent tweets containing hate speech, and the blue points represent tweets that do not contain hate speech. The left pane shows a simple linearly separate classification problem. The right-hand pane shows a more complicated non-linear classification problem.

Method. For the first iteration of the experiment, I used the Chronicling America vectors that were derived from the 1780s. I built a vector representation of each tweet by summing the vector representing each word in a tweet and diving by the number of words in the tweet. The dataset of tweets was split into a training set (n = 4194) and a test set (n = 1399).

I wrote a program in R that split the 4194 tweet training dataset into 10 roughly even groups (or folds to use the lingo of k fold cross validation). During each of k = 10 iterations, roughly 3775 labelled examples (i.e., the vector representation of the tweets with their associated category) are used to fit the Random Forest model. To produce a balanced dataset, the model

sampled an equal number of tweets belonging to both categories (*hate speech* and *not hate speech*). The model was then evaluated by being presented the remaining roughly 420 unlabelled examples (i.e., the vector representation of the tweets without their associated category) and the model made its classification decision for each of the roughly 420 unlabelled tweet vectors. The percent correct score was recorded. The result of this procedure was 10 accuracy scores that were averaged to produce an unbiased estimate of the model's ability to classify novel tweets to one of the two predefined categories. Finally, the model was tested on the test set of 1399 unlabelled tweets and the percent accuracy was recorded to estimate the true ability of the model to classify novel tweets that were never part of the model training process. Importantly, for this final test the two categories of tweets were not balanced (i.e., many more tweets were in the test dataset to represent the true distribution of tweet categories).

The program iterated though this procedure for all sets of decade vectors from the 1780s to 1960s for the Chronicling America corpus and all sets of decade vectors from the 1810s to 2000s for the COHA corpus. To analyze the effect of the decade the vectors were derived from on performance, I conducted a simple linear regression with the model's classification accuracy as a function of decade for each corpus individually. I evaluated the context vectors and the order vectors. The results for the order vectors for all experiments in this manuscript can be found in Appendix D.

Expected results. I expected several results. I expected that the vectors used for all decades would classify novel tweets statistically better than a chance model (50 percent correct). I also expected a statistically significant relationship between the decade of text that was used to build the representation of the tweet, with more recent vectors producing more accurate classification. The reason I expected this result is because our notion of hate speech and

discriminatory sentiments has evolved greatly overtime, with laws preventing hate speech in many countries being examples of this change of values and attitudes compared to the not-so-distant past.

Finally, I expected the models to perform worse than Davidson et al.'s (2017) models that scored an impressive 90 percent accuracy, for two reasons. First, the vectors I built for each decade do not have word vectors for many of the common words people may use in tweets collected in the 21st century (e.g., *selfie, lol*, and emojis such as *:D*). Thus, it is likely that a lot of important information was missing from my representation of the tweets. Second, all the vectors I built were from newspapers (in the case of the Chronicling America corpus) or a combination of newspaper, fiction, non-fiction, and magazines (in the case of the COHA corpus) rather than social media or even web pages that would likely yield a better representation of word meaning given the task. However, it was not my purpose of this experiment to produce classification results as accurate as Davidson et al. Rather, it was my purpose to show that the vectors I have derived capture meaning and can be used to detect a social concept like hate speech using a large and validated database, and to show that the performance of the vectors I derive depend upon the decade of the text they were derived from.

Results. Figure 4.2 shows the results from Experiment 4 using the training data. The plot shows the model's accuracy at detecting hate speech (as measured by percent correct) as a function of the decade of the corpus the vectors were derived from. Each point represents the model's performance for a given decade with the points colored according to the corpus they were derived from. The solid line shows the least squares regression line.

There are several results to note. First, the vectors of every decade performed significantly better than that would be expected by chance (e.g., a 50 percent chance model since

there are two equally distributed classes). Second, though the least squares regression line is visually sloping upwards, there is no statistically significant relationship between the decade and the proportion correct ($R^2 = 0.04$, $p = 0.416$ for the Chronicling America corpus and $R^2 = 0.09$, $p = 0.188$ for the COHA corpus).

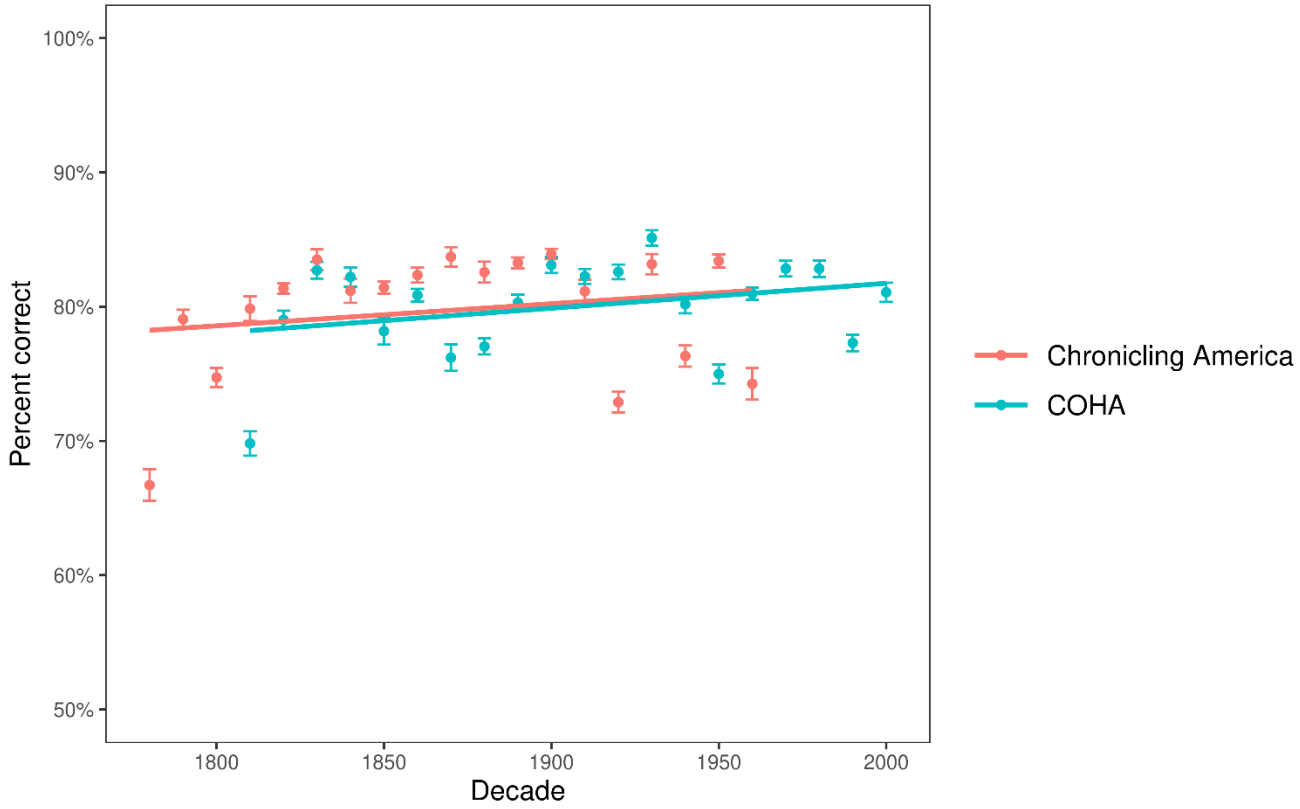


Figure 4.2 Results from experiment 4 (using the cross validated training data). The x axis displays the decade and the y axis displays the percent correct for the Twitter hate speech classification task. Each point represents the accuracy for a given decade and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade. Error bars show the standard error of the mean.

Figure 4.3 shows the results from experiment 4 using the test data. These data contained tweets that were only used for the purposes of evaluating and testing the model and were not used for training. If the model performs well only on the training data and not the test data, it can be an indication that the model is memorizing patterns in the data that do not generalize to novel data (i.e., overfitting). The plot shows the model's accuracy at detecting hate speech (as measured by percent correct) as a function of the decade of the corpus the vectors were derived from. Each point represents the model's performance for a given decade with the points colored according to the corpus they were derived from. The solid line shows the least squares regression line.

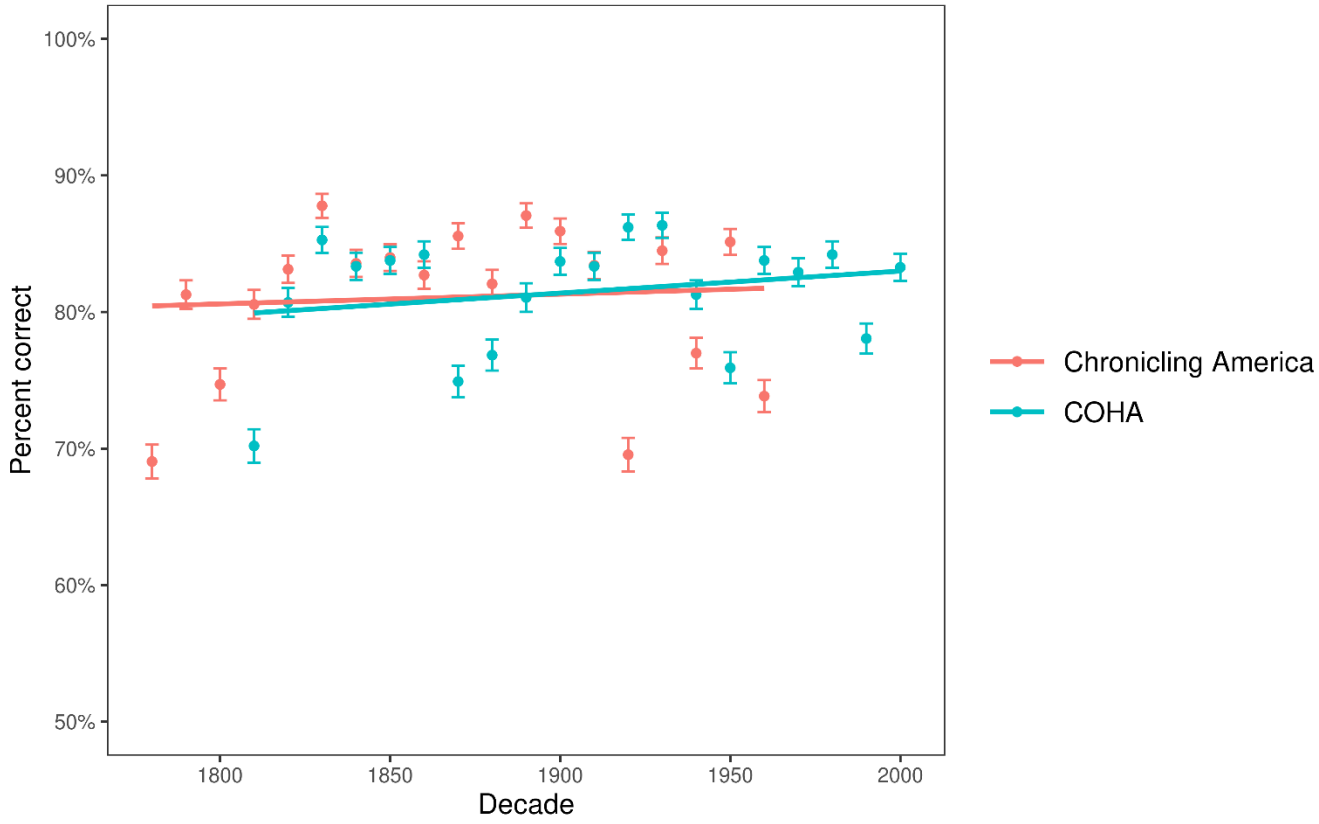


Figure 4.3 Results from experiment 4 (using the test data). The x axis displays the decade and the y axis displays the percent correct for the Twitter hate speech classification task. Each

point represents the accuracy for a given decade and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade. Critically, the accuracy reported in this plot is computed using data the model was not trained on. The error bars show the standard error of a proportion.

There are several results to note. First, the vectors of every decade performed significantly better than that would be expected by chance. Second, though the least squares regression line is visually sloping upwards, there is no statistically significant relationship between the decade and the proportion correct (R^2 = 0.01, p = 0.770 for the Chronicling America corpus and R^2 = 0.05, p = 0.338 for the COHA corpus). In both the training and testing data, the model performed appreciably better than chance, yet the performance did not increase over time. One potential reason for this is that maybe the words that are characteristic of hate speech do not dramatically change over time. For example, racial slurs, or other biased and bigoted words may have remained relatively similar over time.

As expected, the vectors I have derived performed well with four validation tasks: the TOEFL, matching human judgements in word similarity, word classification, and hate speech classification. This final result shows that the vectors are also able to perform in a real-world task of classifying hate speech.

Experiment 5: Classifying Gender Bias in Hiring Decisions

Past research using both the frequency-based approach (e.g., Johns and Dye, 2019) as well as semantic approaches (e.g., Caliskan et al., 2017) has found systematic biases in text as well as the vectors derived from that text. Johns and Dye (2019) found that across all time spans,

genres, and independent of author gender, males are more frequently represented in books than women. Caliskan et al. (2017) demonstrated that like people, semantically derived vectors from a web corpus contained human-like biases against women, racial minorities, and the elderly.

For this section of this thesis, I wanted to demonstrate that the associations and concepts I measured within the semantic vectors in experiments that follow relate to the world and have real consequences. As an example, research may show that women are portrayed in the news less positively than men or that women are less associated with career than men, but does this have any consequence in the real world? In essence, I want to show that these associations that people have that can be measured with semantic vectors in text analysis, are also associated with their decisions on how to behave in the real world.

Kahneman and Tversky studied complex judgement and decision within a rich theoretical framework (e.g., Kahneman, Slovic, & Tversky, 1982). However, they studied these phenomena through very simple experiments, typically inviting people to make a two-choice forced decision, such as asking people their choice of whether they would accept a medical treatment (yes or no). Their experiments allowed Kahneman and Tversky to make enormous contributions to judgement and decision making in the domain of behavioural economics, and in 2002 Kahneman was awarded a Nobel Prize for his work with Tversky (who passed away in 1996).

I used a similar approach to Kahneman and Tversky to study people's decisions. One decision of important consequence is job hiring, yet the job hiring process is far from bias free. As an example, social psychologists have demonstrated that resumes from people with names like Emily and Greg are more likely to get a call for an interview than are resumes from people with names like Lakisha and Jamal (Bertrand & Mullainathan, 2004). Given the importance of bias free job hiring, I will conclude the thesis by demonstrating that I can measure bias in text

that is associated with real world consequence using the tools and methods I have developed in this thesis. For this study, I am going to focus on measuring and predicting peoples' bias in the real-world consequential decision of hiring.

For this experiment, I collected written text from participants as well as had them make a series of hiring decisions. I measured participants' bias to hire males compared to females (or whether the participant exhibits no measurable bias). I then used the methods I have outlined in this thesis previously to measure participants' bias in their written examples of language. The goal was to show that I could use the methods I have previously described for analyzing text to predict a person's real-world behaviour. Specifically, I used the same methods of representing text as the average of its word vectors and the same classification model (Random Forest) in a two-category classification task (*male hiring bias* and *female hiring bias*).

Method. Nine hundred and eighty-eight participants were recruited from the University of Manitoba Psychology participant pool. I did not collect any demographic information from participants as I had no analysis plans that included sex, gender, age, or any other personally identifiable demographic variables.

Participants were told this was a study investigating peoples' hiring decisions for a management or leadership position. For this experiment, I collected two pieces of data from each participant through an online study that participants completed on a computer they had access to through the online survey software Qualtrics. First, I had each participant write an essay that was a minimum of 250 words describing what they thought makes an ideal candidate for a management or leadership position. Specifically, participants wrote an essay in response to the prompt:

Please describe what characteristics make a good business manager or business leader. We are interested in your opinions and thoughts. Be as specific as possible. You can write about any characteristics you think are relevant (personality, attitudes, hobbies etc). Please write at least 250 words.

Second, I had each participant evaluate a set of 20 short (2-3 sentence) descriptions of potential candidates (see Appendix C). For example:

Noah/Emma *is a recent MBA graduate from the University of Manitoba.* ***He/She*** *takes a hands-on approach to managing employees and describes* ***him/her****self as outgoing and likeable.* ***His/Her*** *hobbies include going out on the weekend with friends and traveling.*

Each description contained either a male or female name and was balanced so that 10 descriptions contained male names and 10 contained female names. I used the 10 most common names for each gender listed on the Social Security Services baby names website. Participants were asked to rate how much they endorse a job candidate on a Likert-type scale from 1 (*do not recommend*) to 7 (*highly recommend*). In the end, I had 988 essays and 10 ratings of female job candidates and 10 ratings of male job candidates.

To analyze the data, I wrote a program in R, that computed each participant's average ranking for male candidates ($\overline{x}_{male}$) and their average rating for female candidates ($\overline{x}_{female}$). For each participant I computed a difference score with the average female candidate ranking subtracted form male candidate ranking ($\overline{x}_{male} - \overline{x}_{female}$). Lower scores indicate bias towards hiring female candidates and positive scores indicate bias towards hiring male candidates. Figure 4.4 shows the distribution of difference scores.

To form two groups for the classification task, I selected the top 100 lowest scores (participants who exhibited the most bias to hire female candidates) and the 100 highest scores

(participants who exhibited the most bias to hire male candidates). This process results in a balanced dataset for a two-category classification task, one group of participants who exhibited female hiring bias participants and one group of participants who exhibited male hiring bias.

Like Experiment 4 where I built a model to detect hate speech, I used semantic models and machine learning classification models to build a representation of each participant's essay by averaging the word vectors for each word in their essay. The result of this procedure is a 3000 dimensional vector space, with each of the 200 points in the space representing an essay, with each essay belonging to one category (*female hiring bias* or *male hiring bias*).

As with all previous classification experiments, I used a standard machine learning model (Random Forest) and evaluated the performance of this model on its ability to classify novel essays through the *k* fold cross-validation sampling method described previously. The model's performance was compared to a chance model of 50 percent because there was an equal number of male and female biased participants in the final dataset.

For the first iteration of the experiment, I used the Chronicling America vectors that were derived from the 1780s. I built a vector representation of each essay by summing the vector representing each word in an essay and dividing by the number of words in the essay.

I wrote a program in R that split the 200 essays training dataset into 10 roughly even groups (or folds to use the lingo of *k* fold cross validation). During each of $k = 10$ iterations, 180 labelled examples (i.e., the vector representation of the essays with their associated category) are used to fit the Random Forest model. The model was then evaluated by being presented the remaining 20 unlabelled examples (i.e., the vector representation of the essays without their associated category) and the model makes it classification decision for each of the 20 unlabelled essays vectors. The percent correct score was recorded. The result of this procedure were 10

accuracy scores that were averaged to produce an unbiased estimate of the model's ability to classify novel essays to one of the two predefined categories.

The program iterated though this procedure for all sets of decade vectors from the 1780s to 1960s for the Chronicling America corpus and all sets of decade vectors from the 1810s to 2000s for the COHA corpus. To analyze the effect of the decade the vectors were derived from on performance, I conducted a simple linear regression with the model's classification accuracy as a function of decade for each corpus individually. I evaluated the context vectors and the order vectors. The results for the order vectors for all experiments in this manuscript can be found in Appendix D.

Expected Results. I expected that the models will perform appreciably better than chance, indicating that there is a relationship between the language people use to write an essay and their behaviour in a consequential hiring decision that is detectable using the methods in this thesis. For completeness, as with all previous experiments I built a different set of essay vectors for each decade using both the Chronicling America and COHA corpora.

Results. The average difference score across all 988 participants was -0.036 with a standard deviation of 0.31. Upon visual inspection of Figure 4.4, participants did not strongly show more bias towards males or females on the whole given the distribution of difference scores is fairly symmetrical and the median and mean are very close (0 and 0.31 respectively). However, there was considerable variation and range (*min* = -1.5, *max* = 1.67). Figure 4.4 show the distribution of difference scores.

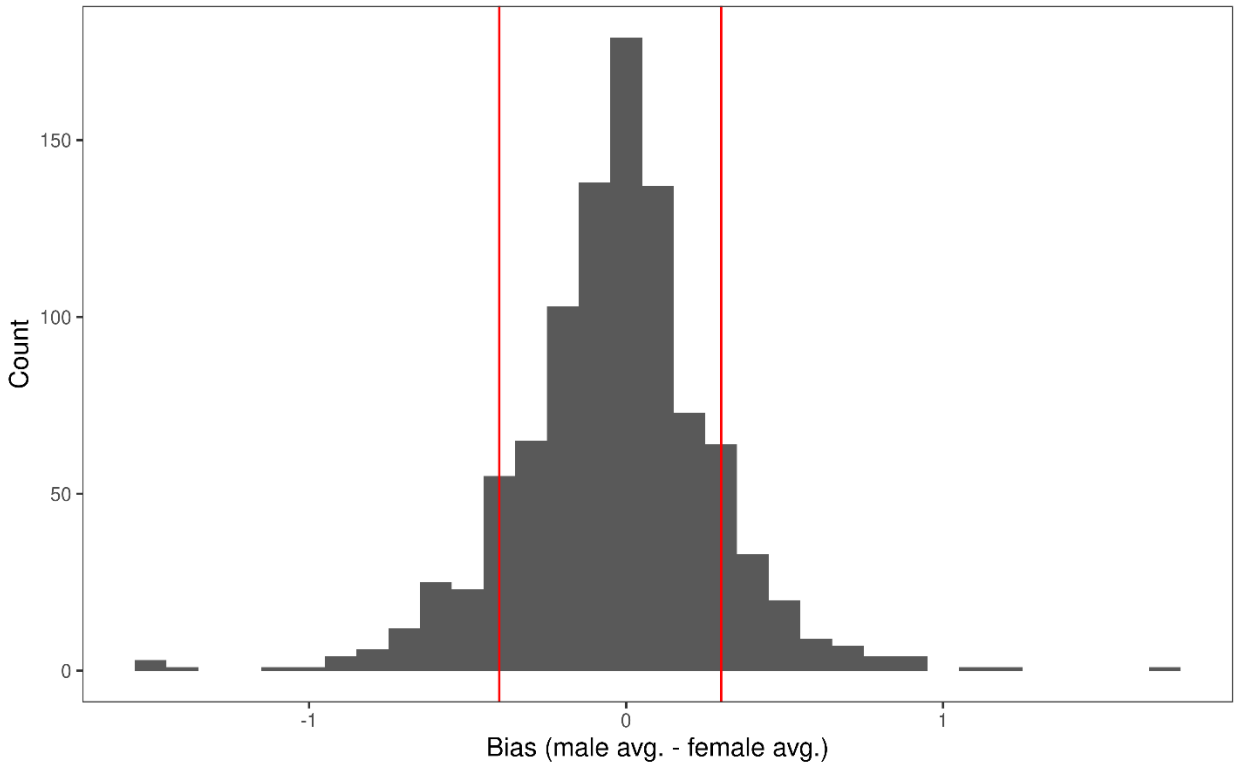


Figure 4.4 Distribution of difference scores for experiment 5. Difference scores are computed as the average ranking of the 10 female candidates subtracted from the average ranking of the 10 male candidates. Negative scores indicate participants a female hiring bias and positive scores indicate participants with a male hiring bias. The red vertical lines show the 10th and 90th percentiles. Participants with a difference score less than the 10th percentile composed the female hiring bias group and participants with a difference score greater than the 90th percentile composed the male hiring bias group.

Figure 4.5 shows the accuracy of the Random Forest model with the vectors for all decades from 1780s to the 2000s using both the Chronicling America corpus and the COHA corpus.

There are several results to note. First, the model performs better than the expected chance model (i.e., 50% correct) across all decades with both corpora. Second, even at its best performance (64.5%), the model does not perform as well as hoped. Third, the relationship between decade and percent correct is not significantly different than 0 ($R^2 = 0.09$, $p = 0.203$ for the Chronicling America corpus; $R^2 = 0.03$, $p = 0.437$ for the COHA corpus).

The main reason I think the model did not perform well is because of the data collected. In many machine learning tasks, the dependent variable (i.e., bias towards males or bias towards females) would be annotated by a third party. Typically, crowd sourced participants would read through the essay and provide labels based on their reading of the essays. Using this approach to data labelling some detectable relationship is forced between essays and the categories attempted to be predicted by the model. For example, annotators might read an essay and deem it to exhibit bias against female candidate due to a sexist remark or other subtle or not-so-subtle meaning. However, in my study, there was no forced correspondence between the categories of essay and the content of the essays. Participants who wrote the essay might not exhibit any measurable bias in their writing that matches their behaviour in the task of providing rankings on candidates.

In retrospect, this may have been too subtle of an effect to detect reliably and strongly. There did not seem to be a strong correspondence between participants' essays and their behaviour in the hiring decisions. Even though the results of this experiment were not as promising as I had hoped, I feel confident moving forward due to the models' performance on previous more structured tasks.

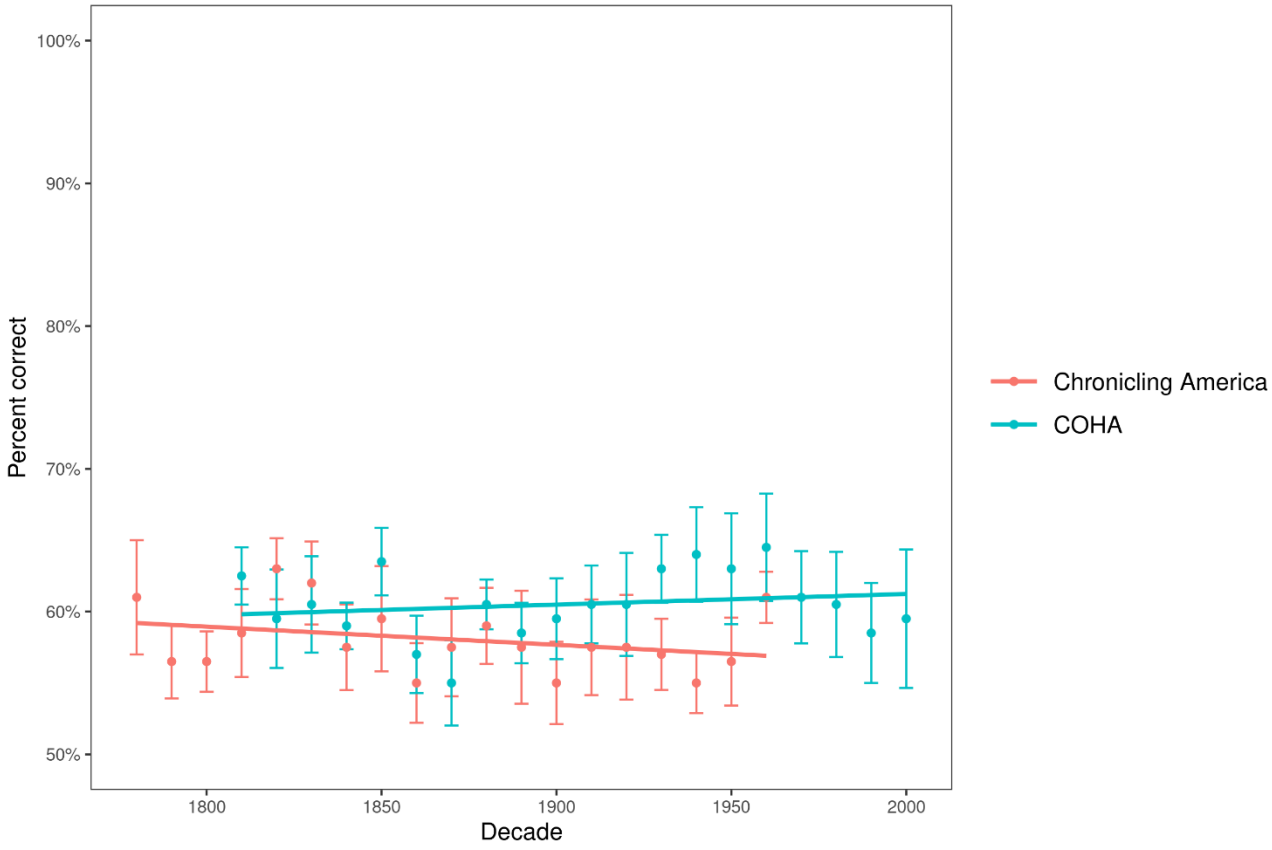


Figure 4.5 Results from experiment 4. The x axis displays the decade and the y axis displays the percent correct for the hiring bias classification task. Each point represents the accuracy for a given decade and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade. Error bars show the standard error of the mean.

Even though there was no overall pattern of bias found in participants (i.e., the average bias was very close to zero), it is possible that there are subgroups within my participants that are strongly biased towards males or females. Given that I did not have demographic variables collected from participants, I could not conduct follow up analyses to see if the pattern of bias (i.e., male versus female hiring bias) was related to age, gender, or other demographic variables.

Had I collected these variables it would have been interesting to determine which demographic variables could account for differences in bias and potentially build a model on the smaller subset of data (i.e., only use data from men who were biased to hire men).

Interim Summary

Thus far, the five sets of experiments I have presented give an indication of the meaning contained in the vectors I derived. Importantly, these experiments tested the vectors I derived with a specific performance measure (e.g., cosine similarity or percent correct) using a validated ground truth. I now want to move from using the model in simple, somewhat artificial tasks, such as synonym tests, to using the model to measure social concepts, such as stereotypes, throughout time. For measuring social change and differences in concepts, I will no longer have a specific performance measure, but rather I will only be able to report the changes and differences I observed.

Chapter 5: Investigating Real World Meaning

Investigating Vector Meaning

In the experiments that follow, I investigated several areas of society ranging from gender stereotypes to war and developed methods for assessing how these concepts changed from the late 1700s to the present day. Because the concepts I was interested in studying could not be studied directly, I used certain words that stand in for concepts. By doing so, I hoped to uncover change in a concept's meaning through time.

Experiment 6: Measuring Concept Valence

One standard method of investigating the meaning of particular word vectors is the nearest neighbors approach. In this approach, the vector for a word of interest, such as *computer*, is compared to all other word vectors and the top *n* words and their cosine similarity values are reported. As an example, Jones and Mewhort (2007) reported that the top five nearest neighbors for the word *computer* were *data*, *computers*, *processing*, *processed*, and *storage*. Like a dictionary that explains a word's definition by reference to other words, the nearest neighbors approach shows the meaning of a word by referencing it to the other most similar words in the corpus.

The advantage of this approach is obvious: by looking at which words are most similar to a target word, it allows for a quick investigation of the meaning of any given word and allows one to validate the meaning of the vectors. However, given that I wanted to investigate many words and relationships over many decades, this approach would become unmanageable. Furthermore, this approach starts to veer into the realm of qualitative analysis, and especially for analyzing contentious social concepts such as stereotypes, I tried to stay as quantitatively minded as possible to avoid too much subjectivity from creeping into my analyses. It is too easy and too

tempting to shoehorn the results of the nearest neighbor approach into a post-hoc justification. This becomes a problem in many types of quantitative analyses such as naming factors in factor analysis, or in naming topics in topic modelling. My approach in this experiment was to find a more objective measurement of elements of meaning that can be used to evaluate any given concept and the change in that concept over time.

I retained the strengths of the nearest neighbor approach for investigating word meaning. However, I solved the problems I just described by simplifying the analysis from analysing the exact words found in the list of the nearest neighbors and trying to judge whether they make sense and have changed over time, to looking at the sentiment or emotionality (e.g., positivity/negativity) of those words. Researchers have decomposed emotionality of a word into three main factors which are the valence (pleasantness of a word), arousal (intensity of a word), and dominance (degree of control of a word; Warriner, et al., 2013). Analyzing the emotionality of words through the word's valence, dominance, and arousal of the nearest neighbors of the concepts provides understanding of whether the concepts in question are viewed positively or negatively, dull or intense, and controllable or uncontrollable relative to other concepts or the same concept throughout time. Cognitive Psychologists have compiled large databases where participants rate the qualities of words such as valence, concreteness, and imageability (e.g., Bradley & Lang, 1999; Warriner et al., 2013). But given that I was interested in the emotionality of words, I made use of Warriner et al.'s (2013) collection of almost 14,000 word ratings of valence, arousal, and dominance. These words were rated on a 9-point Likert-type scale with higher scores indicated more positive feelings (valence), greater intensity (arousal), and more control (dominance). The 9-point scale ranged from unhappy to happy (valence), calm to excited

(arousal), and controlled to in control (dominance)[2]. Participants were told that the lower end of the valence scale represents feeling unhappy, annoyed, unsatisfied, melancholic, despaired, or bored, and the upper end of the valence scale represents feeling pleased, satisfied, contented, hopeful; the lower end of the arousal scale represents feeling relaxed, calm, sluggish, dull, sleepy, or unaroused, and the upper end of the arousal scale represents feeling excited, frenzied, jittery, wide-awake, or aroused; the lower end of dominance scale represents feeling controlled, influenced, cared-for, awed, submissive, or guided, and the upper end of the dominance scale represents feeling in control, influential, important, dominant, autonomous, or controlling. As an example of how this method could be used, one could look at the mean valence rating for the top 10 nearest neighbors for both the words *music* and *weapon*. We would expect (and find) that the words associated with *music* are more positive than the words associated with *weapon* and that this relationship does not change over time.

Method. For the first iteration of the experiment, I used the Chronicling America vectors that were derived from the 1780s. I found the top 100 most similar words to the word *war* by computing the cosine similarity between *war* and every word in the 1780s Chronicling America corpus. I computed the average emotionality of the top 100 words as measured by valence, arousal, and dominance by using the Warriner et al. (2013) norms. The program iterated though this procedure for all sets of decade vectors from the 1780s to 1960s for the Chronicling America corpus and all sets of decade vectors from the 1810s to 2000s for the COHA corpus. I evaluated the context vectors and the order vectors. The results for the order vectors for all experiments in this manuscript can be found in Appendix D. Finally, I standardized the measurements of the average valence, dominance, and arousal for each corpora by computing the z scores $((x - \bar{x}) /$

[2] Originally Warriner et al. had the 9-point scale range from happy = 1 and unhappy = 9 for valence, and excited = 1 and calm = 9 for arousal, but reverse coded these two measures to be more intuitive.

sd(x)) of each of these measures to offer a standardized comparison against each other and other empirical data used later in this experiment.

Expected results. For each decade, I expected the words most associated with war to be less positive throughout periods of known warfare (e.g., the 1910s during World War 1 and during the 1940s for World War 2).

Results. Figure 5.1, 5.2, and 5.3 shows the results from experiment 6 for the valence, arousal, and dominance measurements. The x axis shows the decade the vectors were derived from and on the y axis the emotionality of the word *war*. The solid line in each plot show the mean valence, arousal, and dominance for the 100 nearest neighbors in the semantic space to the word *war*. The horizontal dotted line represents the z score corresponding to the center of the valence, arousal, and dominance scale (e.g., a 5 on the 9-point Likert-type scale). The shaded regions of the plot correspond to several of the largest wars during this time period (Civil War (1861-1865), World War 1 (1914-1918), World War 2 (1939-1945), the Korean War (1950-1953), and the Vietnam War (1955-1975)). The plot contains two subplots, one for each corpus. I will address the other measurements shortly.

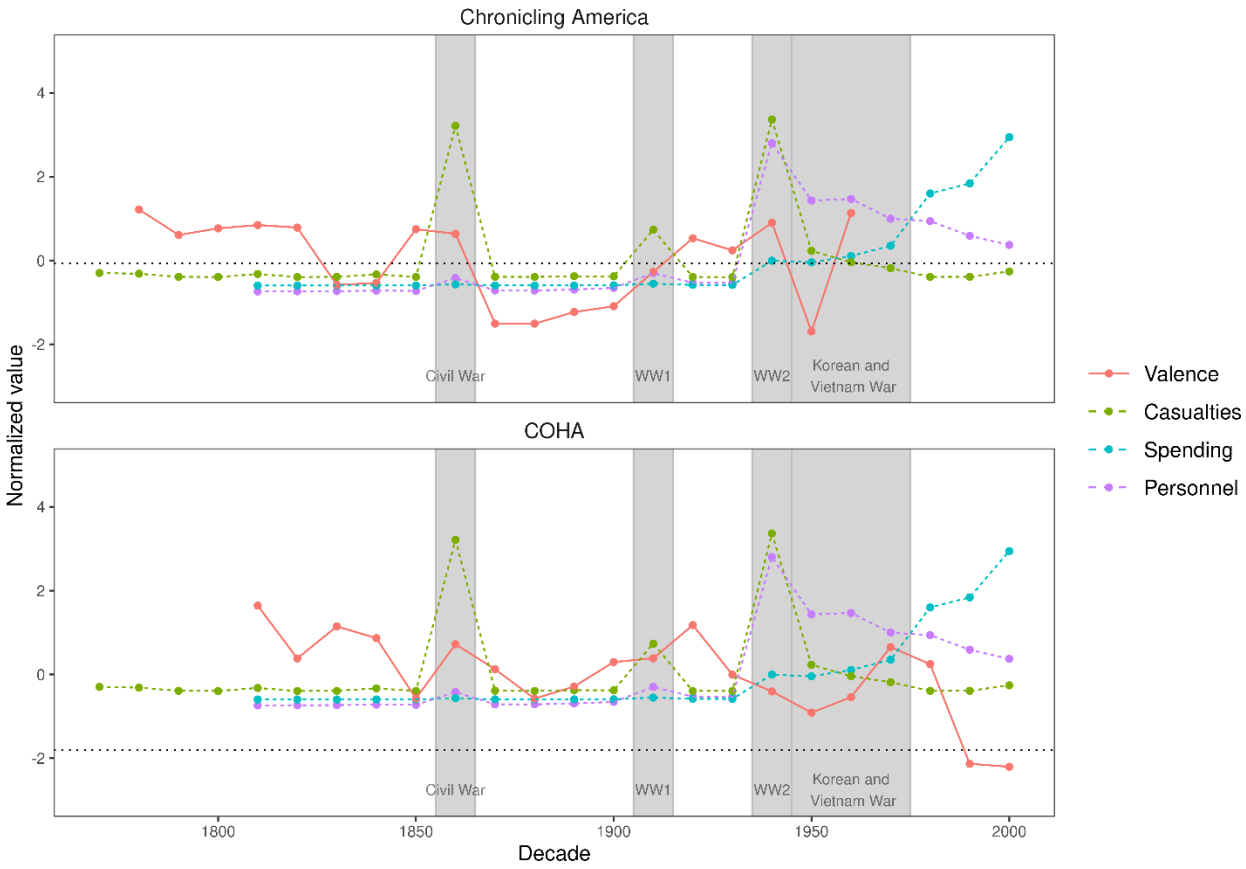


Figure 5.1 Results from experiment 6 for valence. In the two subplots, the x axis displays the decade and the y axis displays the normalized measurement of valence and normalized empirical measurements (war casualties, spending, and personnel). The top subplot displays the results for the Chronicling America corpus and the bottom subplot displays the results for the COHA corpus. All variables are standardized to offer easy visual comparison (i.e., all four variables are converted to z scores). The horizontal dotted line represents the z score corresponding to the center of the valence scale (e.g., a 5 on the 9-point Likert-type scale).

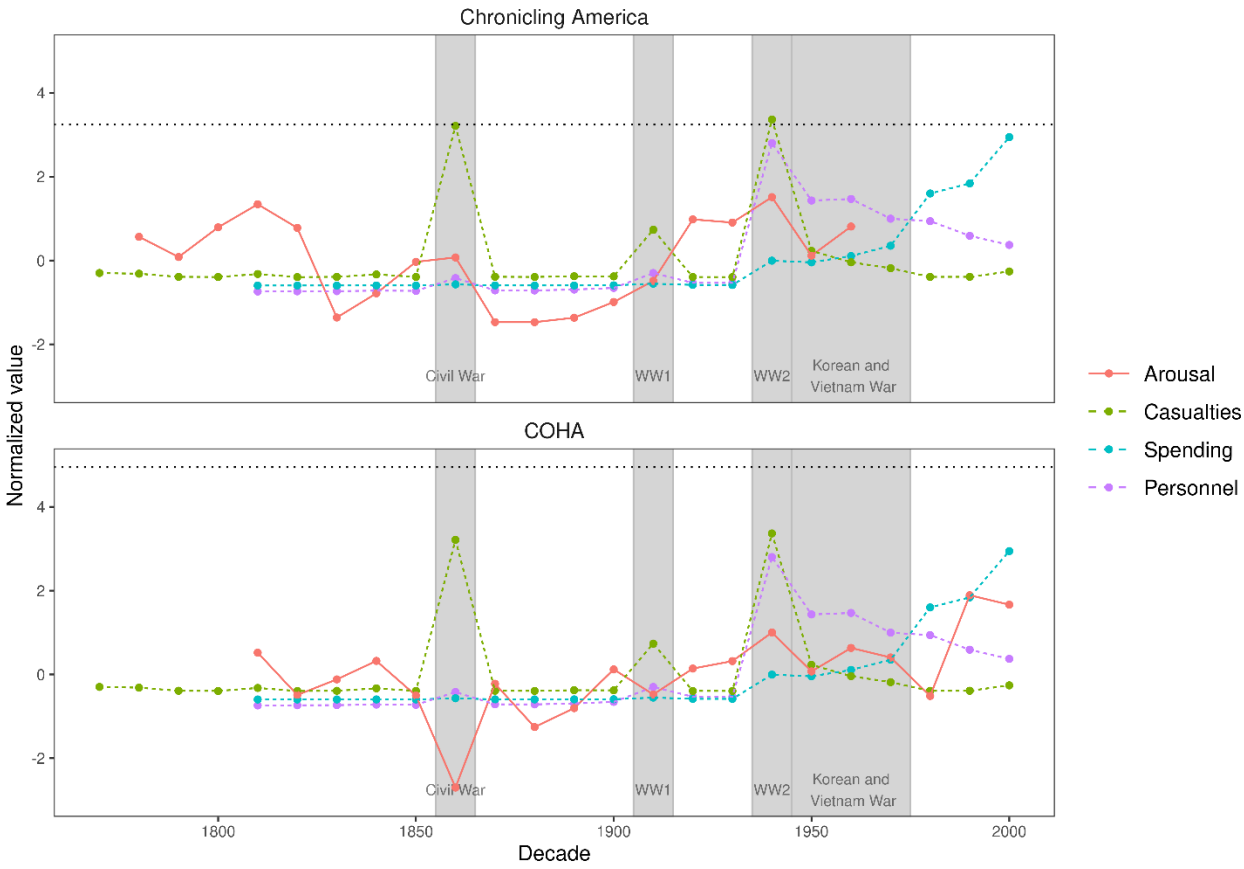


Figure 5.2 Results from experiment 6 for arousal. In the two subplots, the x axis displays the decade and the y axis displays the normalized measurement of arousal and normalized empirical measurements (war casualties, spending, and personnel). The top subplot displays the results for the Chronicling America corpus and the bottom subplot displays the results for the COHA corpus. All variables are standardized to offer easy visual comparison (i.e., all four variables are converted to z scores). The horizontal dotted line represents the z score corresponding to the center of the arousal scale (e.g., a 5 on the 9-point Likert-type scale).

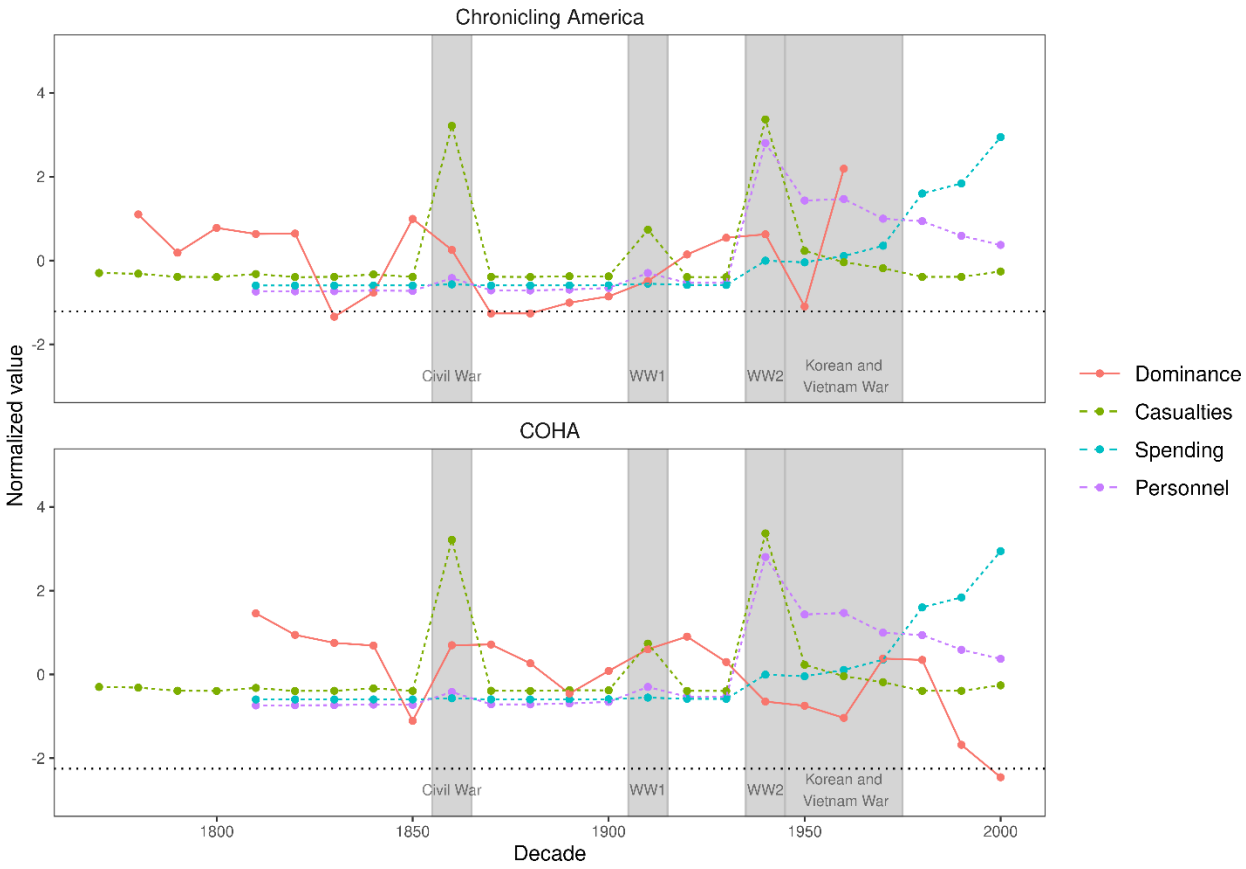


Figure 5.3 Results from experiment 6 for dominance. In the two subplots, the x axis displays the decade and the y axis displays the normalized measurement of dominance and normalized empirical measurements (war casualties, spending, and personnel). The top subplot displays the results for the Chronicling America corpus and the bottom subplot displays the results for the COHA corpus. All variables are standardized to offer easy visual comparison (i.e., all four variables are converted to z scores). The horizontal dotted line represents the z score corresponding to the center of the dominance scale (e.g., a 5 on the 9-point Likert-type scale).

There are several patterns to note. First, the trajectory of each of the three measures are highly correlated. The second result is that the emotionality of the word *war* does not increase or decrease in a linear and straightforward way throughout the decades. This is probably to be

expected given that since the 1780s the United States has been involved with many wars in between periods of peace. However, visually there are three distinctive peaks in the early 1800s, 1850s, and during the first half of the 1900s. These periods correspond to several of the largest and most well-known wars in United States history: War of 1812 (1812-1815), the Mexican-American War (1846-1848), the Civil War (1861-1865), World War 1 (1914-1918), World War 2 (1939-1945), the Korean War (1950-1953), and the Vietnam War (1955-1975).

In the Chronicling America corpus there are some interesting results to note. Valence drops quite sharply after the civil war, and slowly starts increasing until after World War 2 where it drops sharply again. These two drops in valence are right after a large number of casualties in the Civil War and World War 2. It is possible that the valence measurement is picking up on the increase in despair and grief following such a large loss of lives in these two massive wars.

All three measurements in the Chronicling America corpus also increase quickly after the sharp decrease after World War 2 compared to the length of time it takes for the measurements to rebound after the drop after the Civil War. One reason for this pattern might be because the World War 2 and the Korean and Vietnam wars were abroad and didn't have the immediacy of the Civil War which was happening on American soil.

The measurements for arousal in the Chronicling America all fall below the average valence ratings across all words collected by Warriner et al. This is surprising giving that war would typically be thought as high in arousal. One reason for this finding could be that the language used when discussing war often tends to mask the true meaning and appropriate arousal, such as the term *casualties* as opposed to *death*.

The valence, arousal, and dominance measurements also have some interesting patterns for the COHA corpus. The pattern of results across time is very similar across the three

measurements. One notable difference in the measurements is that whereas valence and dominance drop to the lowest observed levels in the late 1900s, arousal increases to its highest observed level. This might signify our changes attitudes towards war — that is a terrible (i.e., low valence), it arouses fear and despair (i.e., high arousal), and that once started it is uncontrollable (i.e., low dominance).

However, this explanation of the pattern of emotionality over 200 years is based on a visual inspection of the data. What I wanted to do is find several sources of empirical data related to the United States war involvement to evaluate how well those empirical data fit with the model's judgements on the emotionality of the word *war*. If the emotionality of war changes with these empirical sources of data, that would give some further credibility that these changes in emotionality are capturing changes in the meaning of war.

I used three sources of empirical data. First, I used data on the total number of U.S. war causalities. Presumably, the meaning of war in newspaper over the centuries may depend on how many soldiers are involved in wars at the time and how many people were killed as a result of that war.

My second and third sources of empirical data come from *Our World in Data*. Specifically, I used their data on Military expenditures per capita from 1816 to 1970. I also used their data on the number of Military personnel from 1816 to 1970 which defines military personnel as the number of troops ready for combat under command of the national government.

The addition dotted lines in figure 5.1, 5.2, and 5.3 shows the three sources of empirical data: the number of U.S. war casualties, military expenditure per capita, and military personnel.

There are several results to note. First, at least upon visual inspection, the number of U.S. causalities, military expenditure, and military personnel all appear to follow the pattern of emotionality to some extent and much more so with the Chronicling America corpus.

One puzzle in the results is why the valence for the word war would be higher during periods of high number of war causalities, spending, and large number of military personnel. At first blush, it would seem a negative relationship (rather than positive) between valence and these empirical measures would be expected. One explanation might be that it is easy to be against war (i.e., low emotionality or valence) during periods of relative peace. In contrast, when politicians or media feel the need to "rally the troops" the benefits of war might be emphasized more, such as fighting and protecting one's country or values thereby making the valence of the word *war* more positive during periods of war.

This method allows for automated measurement of concept meaning, summarizing the three main components of emotionality for any given word or concept. The method can be used to understand differences in concept meaning over time as I have presented here, or across different corpora (e.g., different newspaper publications). Furthermore and importantly, the measurements of emotionality can be evaluated against empirical data across time.

Experiment 7: Measuring the Strength of Concept Association

The word valence method is a purely data-driven approach to analyzing the word vectors for meaning and their changes in meaning over time. In experiment 6, I used war as a case study to measure changing attitudes about war from the 1780s to 2000s. However, there are certain theoretically driven hypotheses researchers may also want to test using similar methods. In this next set of experiments, I used a method for testing how the associations between certain

concepts change over time, using gender stereotypes catalogued by past research (e.g., Caliskan et al., 2017) as a case study to measure attitudes and stereotypes of social groups.

Method. As previously mentioned, one way of building a representation of a concept is to use words as the fundamental building blocks of the concept. To form a concept of gender (men and women) I collected 500 male names and 500 female names that each stereotypically relate to men and women respectively. I used the most common 500 male and female names from the Social Security Services baby names website. I generated a concept vector by summing male names into one vector that represents maleness. I then generated a concept vector by summing female names into another vector that represents femaleness.

I measured the similarity between the male and female concept vectors to six sets of vectors using cosine similarity as a way of assessing gender stereotypes. The six sets of vectors are 1) male words (grandfather, father, dad, son, brother, uncle, nephew, male, man, men, boy, he, his), 2) female words (grandmother mother, mom, daughter, sister, aunt, niece, female, woman, women, girl, she, her), 3) career words (executive, management, professional, corporation, salary, office, business, career), 4) family words (home, parents, children, family, cousins, marriage, wedding, relatives), 5) math words (math, algebra, geometry, calculus, equations, computation, numbers, addition), and 6) art words (poetry, art, dance, literature, novel, symphony, drama, sculpture).

The result of this procedure is a cosine similarity measurement between the 58 words listed above with both the male and female concept vectors. Finally, I computed a difference measurement that is the cosine similarity between a given word and the female concept vector subtracted from the cosine similarity between a given word and the male concept vector. For example, cos(male_concept, grandfather) – cos(female_concept, grandfather). A negative

difference score indicates a stronger relationship between the female concept vector and the word grandfather and a positive difference score indicated a stronger relationship between the male concept vector and the word grandfather.

The male and female words were word lists I created to capture these two concepts and I think the words that belong to each of them are straightforwardly male and female words. The career, family, math, and arts words are words lists taken from past research on measuring gender bias and stereotypes with word vectors (Caliskan et al., 2017). The math and art words come originally from research on the Implicit Association Test (IAT) showing that participants have an implicit bias to associating female terms with art and male terms with math (Nosek, et al., 2002). Similarly, the career and family words also were originally used in Nosek, et al. (2002) demonstrating that people have an implicit bias to associate female terms with family and male terms with career.

Expected Results. I expected several results. First, I expected the vector of male names to be more similar (as measured by cosine similarity) to the male words (e.g., father, grandfather) than female words (e.g., mother, grandmother). Similarly, I expected the vector of female names to be more similar to the female words than male words. These comparisons mainly act as control conditions. Second, and more interestingly, I expected the male names to be more similar to career words than family words, and female names to be more similar to family words than career words. Third, I expected the male names to be more similar to math words than art words, and female names to be more similar to art words than math words. This hypothesis is consistent with past research on gender stereotypes (e.g., Caliskan et al., 2017).

Results. The left panel of figure 5.4 shows the cosine similarity between the male names vector (*George + Bill + John + ...)* and 58 words for the 1900s from the vectors for the

Chronicling America corpus. The bars corresponding to each word are color coded according to six word categories (male, female, career, family, math, and arts). The right panel of figure 5.4 shows the cosine similarity between the female names vector (*Sally + Anne + Jill + ...)* and the same 58 words. The center panel of figure 5.4 shows the difference between the two sets of bars for each of the 58 words (i.e., the similarity of the female names to the word *grandfather* subtracted from the similarity of the female names to the word *grandfather*). Any bars that are greater than zero represent *maleness* as they correspond to the vector of male names more strongly than female names. Bars that are less than zero represent *femaleness* as they correspond to the vector of female names more strongly than male names. To make the results easier to understand I have plotted black bars that represent the average cosine similarity of each word type for each of the three sub plots.

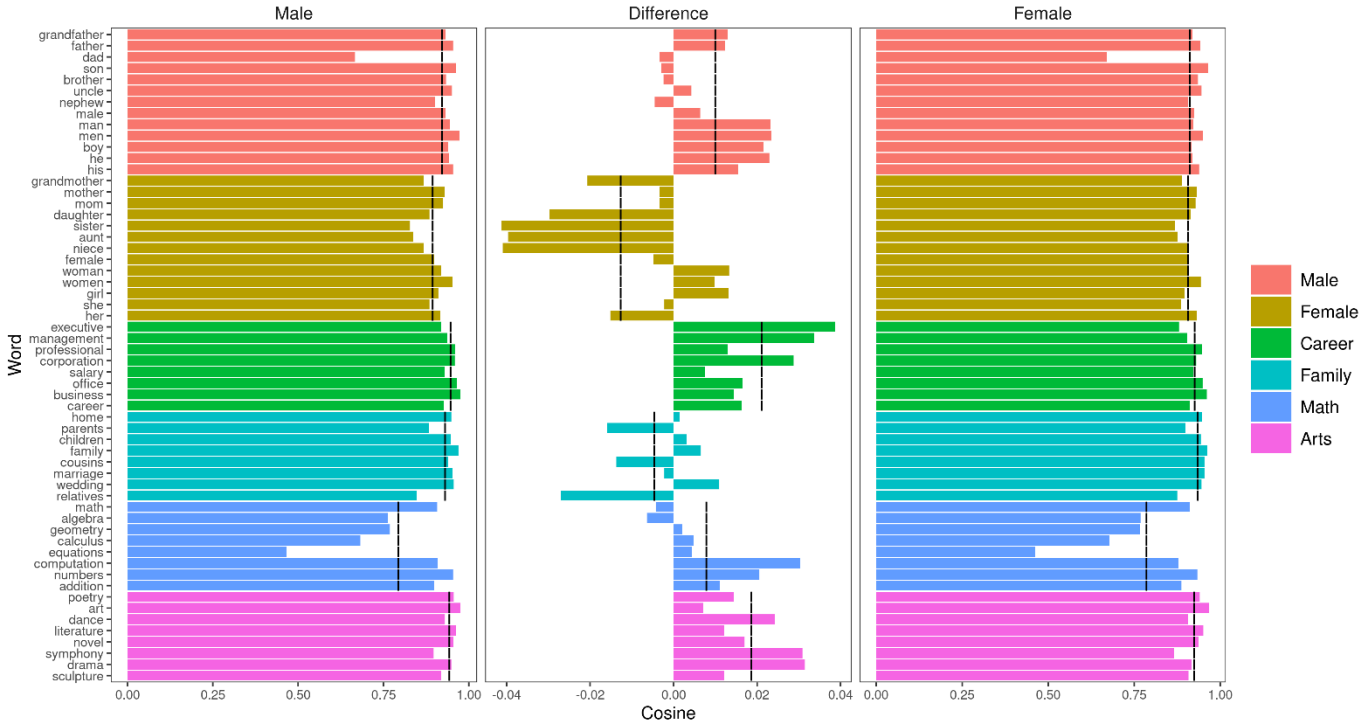


Figure 5.4 Results for experiment 7 (Chronicling America corpus only). In the left pane, the y axis shows 58 words organized according to six color coded categories (male, female,

career, family, math, and arts). The x axis shows the cosine similarity between a male name concept vector and each of the 58 words. In the right pane the y axis shows the same 58 words organized according to six color coded categories (male, female, career, family, math, and arts). The x axis shows the cosine similarity between a female name concept vector and each of the 58 words. The center pane shows the same 58 words as the left and right pane but shows the difference between each pair of bars from the left and right pane (e.g., cosine(male names, grandfather) - cosine(female names, grandfather)). The vertical black lines represent the mean of each set of color bars in a given subplot. The plot displays data from the most recent data of the corpus.

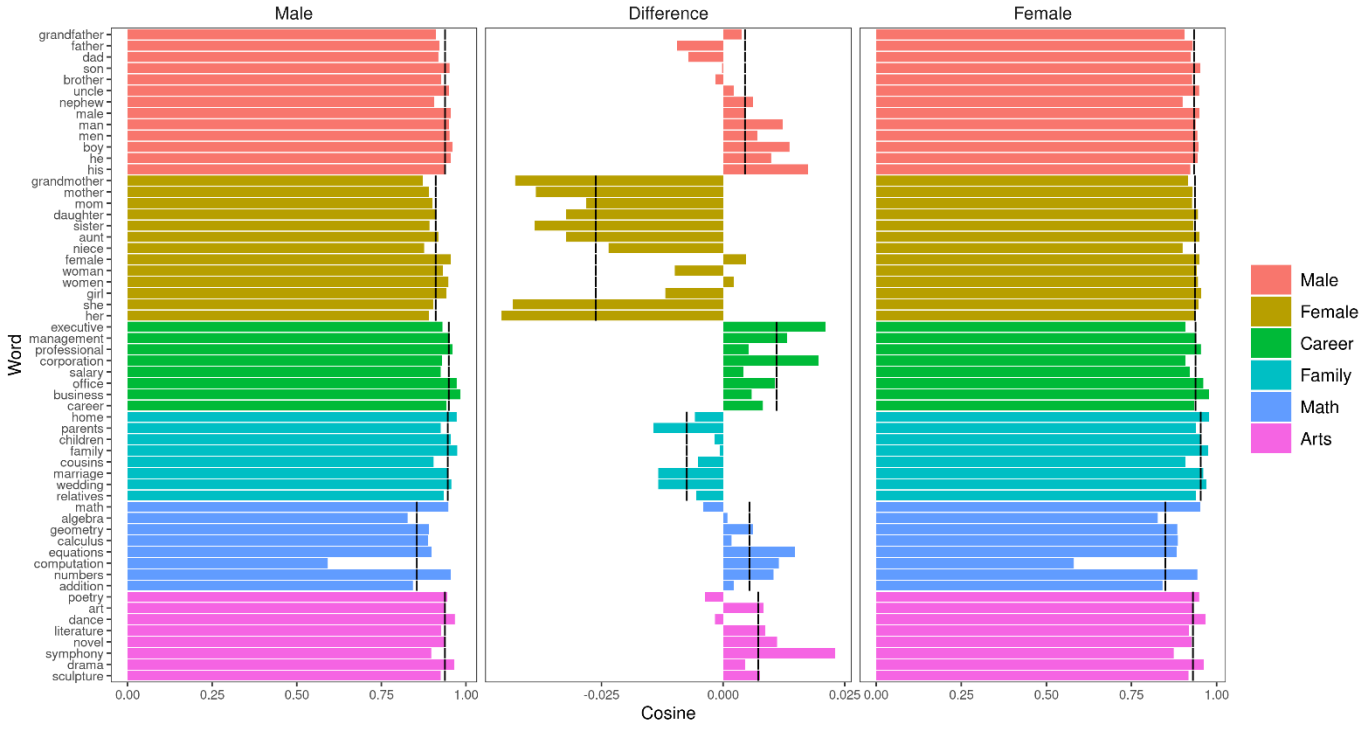


Figure 5.5 Results for experiment 7 (COHA corpus only). In the left pane, the y axis shows 58 words organized according to six color coded categories (male, female, career, family, math, and arts). The x axis shows the cosine similarity between a male name concept vector and

each of the 58 words. In the right pane the y axis shows the same 58 words organized according to six color coded categories (male, female, career, family, math, and arts). The x axis shows the cosine similarity between a female name concept vector and each of the 58 words. The center pane shows the same 58 words as the left and right pane but shows the difference between each pair of bars from the left and right pane (e.g., cosine(male names, grandfather) - cosine(female names, grandfather)). The vertical black lines represent the mean of each set of color bars in a given subplot. The plot displays data from the most recent data of the corpus.

To make the results easier to visualize I have plotted only the center sub plot of figure 5.4 and 5.5 that displays the differences between the male and female concept vectors.

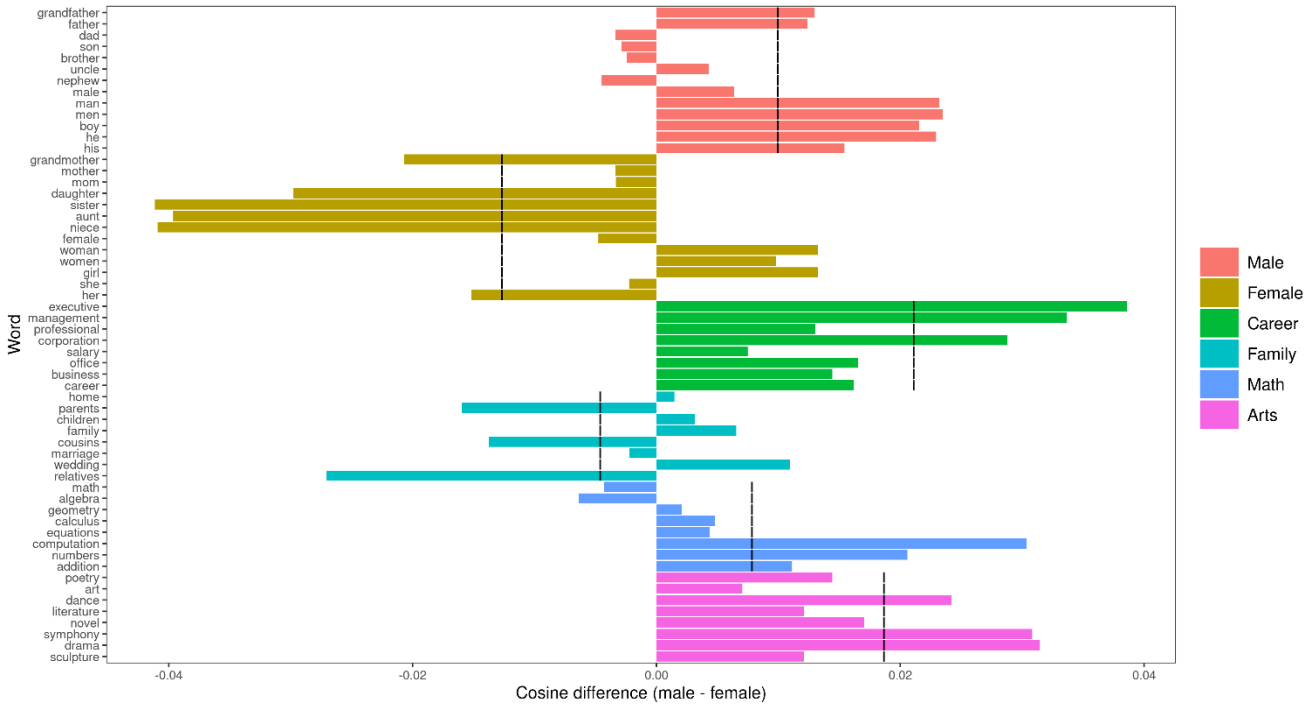


Figure 5.6 Results for experiment 7 (Chronicling America corpus only) focused on the center pane of figure 5.4.

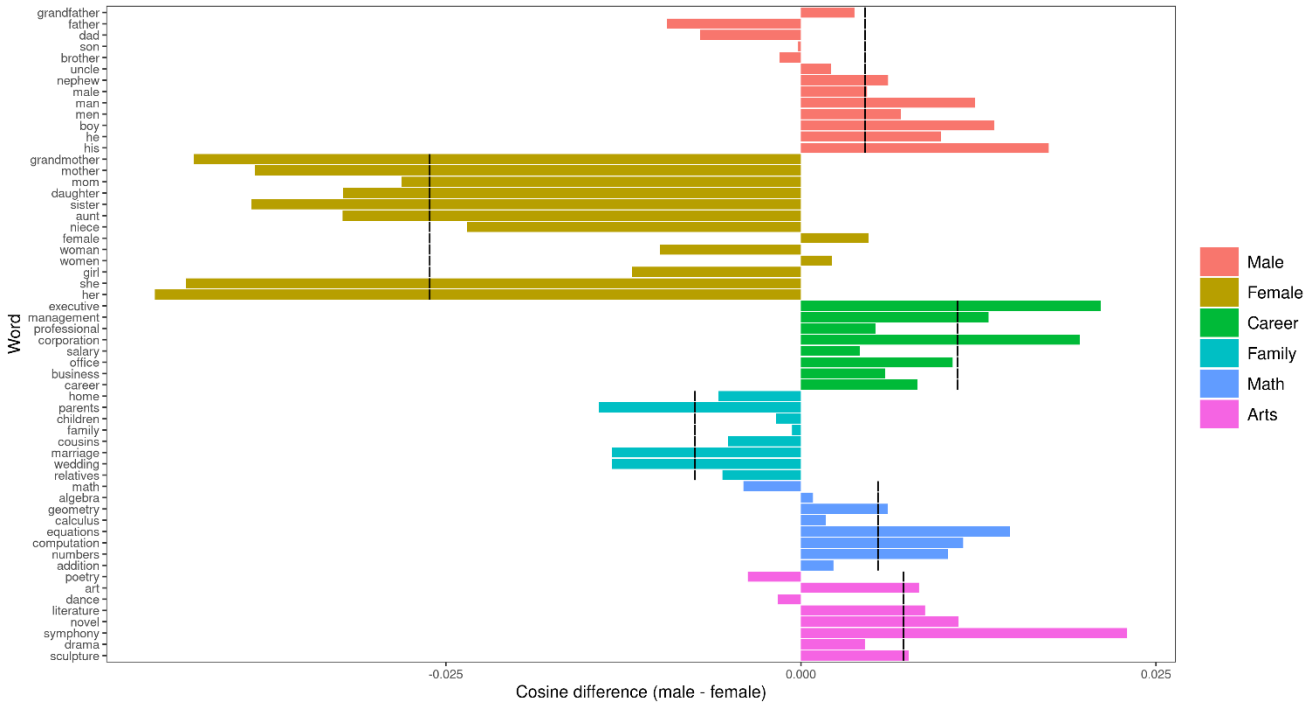


Figure 5.7 Results for experiment 7 (COHA corpus only) focused on the center pane of figure 5.5.

There are several results to note. First, by looking at the average cosine similarity we can largely see the pattern of expected results. The male concept vector relates more strongly to male words than the female concept vector, the female concept vector relates more strongly to the female words than the male concept vector, the male concept vector relates more strongly to the career words than the female concept vector, the female concept vector relates more strongly to the family words than the male concept vector, and the male concept vector relates more strongly to the math words than the female concept vector. However, the one result that was not as expected was that the male concept vector relates more strongly to the arts words than the female concept vector.

Results through time. These results presented have been from the vectors from a single decade from each corpora. Rather than repeating the same plot for each decade, I will plot the six means associated with each category of words for the difference in cosines (the means displayed in figure 5.6 and 5.7) across time. This shows the overall similarity and differences without focusing too strongly on any one particular word comparison.

Figure 5.8 shows the results of experiment 7 for the male and female words. For every decade from 1780 to 1960, the similarity between male names and male words is stronger than the similarity between female names and female concepts for both corpora. This relative ordering shows the expected pattern of results. In addition to the relative differences between these two lines, we can consider the absolute divergence from zero. As a reminder, more positive differences indicate more *maleness* whereas more negative numbers indicate more *femaleness*. Starting in the 1830s the concepts of male and female start to diverge from each other becoming more polarized concepts through to the 1960s. These results are especially pronounced in the COHA corpus where for every decade the male words relate more strongly to the male names than the female names, and the female words relate more strongly to the female names than the male names. The bottom panel of figure 5.8 is also interesting because the polarization between male and female concepts also suggests that the polarization is largely driven by how women are portrayed (i.e., the line representing the similarity between female names and words becomes increasingly female over time).

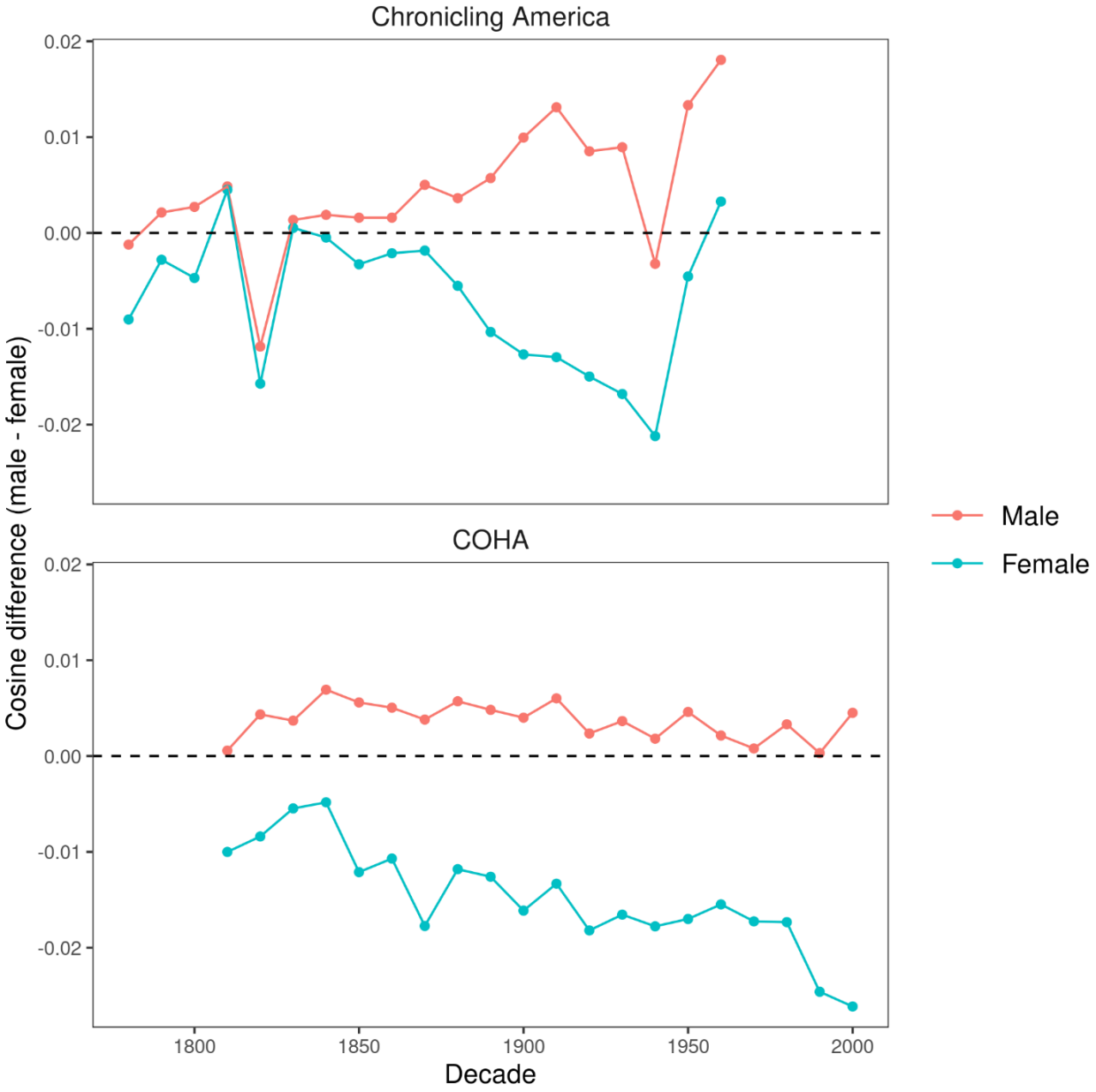


Figure 5.8 Results for experiment 7. The x axis shows decade and the y axis shows the difference in average cosine similarity between the male name concept vector and female name concept vector. The top subplot shows the results for the Chronicling America corpus and the bottom subplot shows the results for the COHA corpus. The points show the average cosine similarity difference color coded by each category of words (male, female).

In figure 5.9 we see the same pattern of results as the previous plot. For every decade from 1780 to 1960 (except the 1820s in the Chronicling America corpus and the 1840s in the COHA corpus), the similarity between male names and career words is stronger than the similarity between female names and family words. This relative ordering shows the expected pattern of results. In addition to the relative differences between these two lines, we can consider the absolute divergence from zero. Starting in the 1830s the concepts of male and female start to diverge from each other becoming more polarized concepts through to the 1960s. In the COHA corpus, there is also an interesting pattern, where starting in the 1920s there is a reversal of the increasing male trend in both career and family words. Many historical events may drive these changes, such as the great depression and differing opinion and attitudes on traditional gender roles.

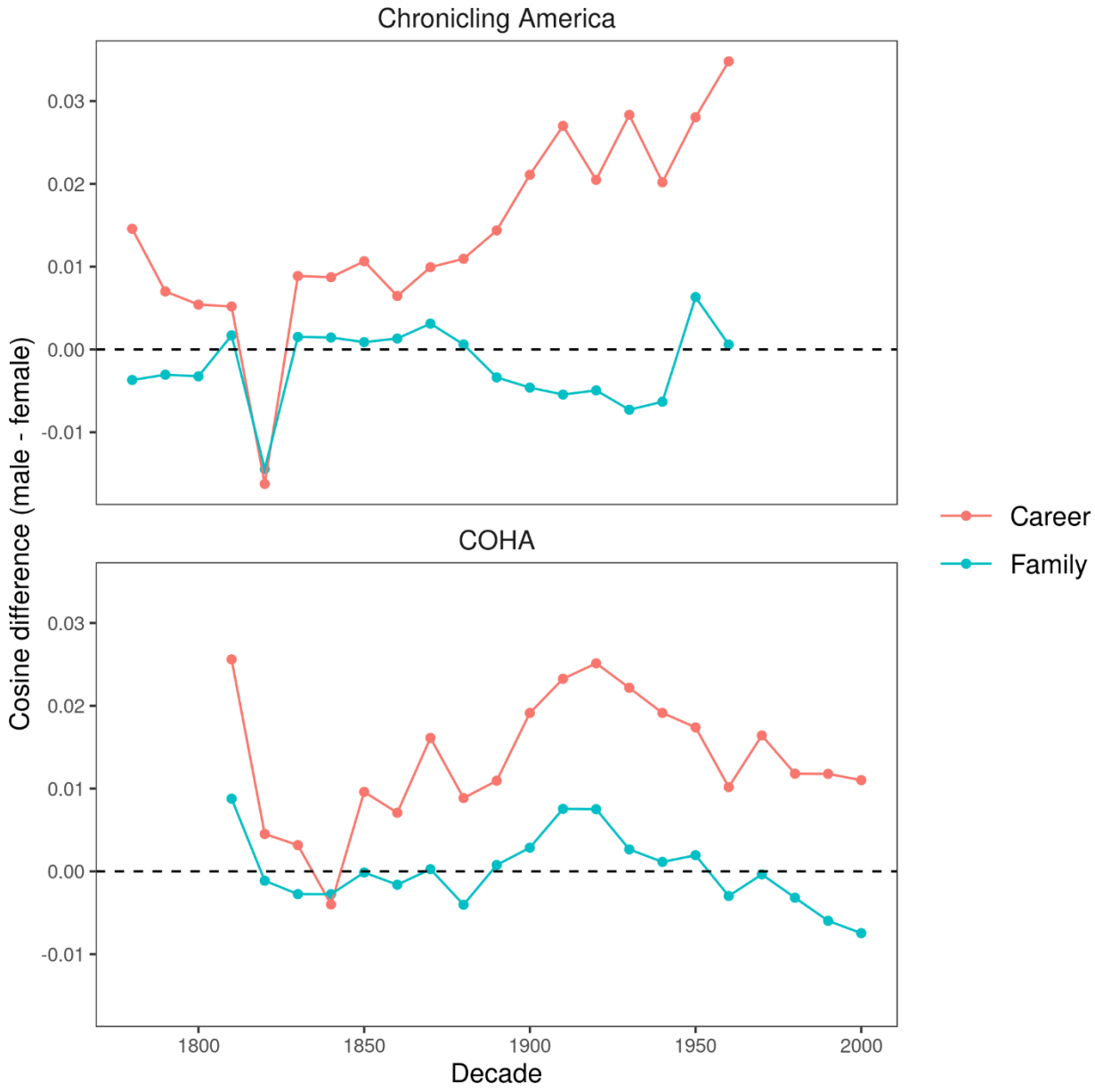


Figure 5.9 Results for experiment 7. The x axis shows decade and the y axis shows the difference in average cosine similarity between the male name concept vector and female name concept vector. The top subplot shows the results for the Chronicling America corpus and the bottom subplot shows the results for the COHA corpus. The points show the average cosine similarity difference color coded by each category of words (career, family).

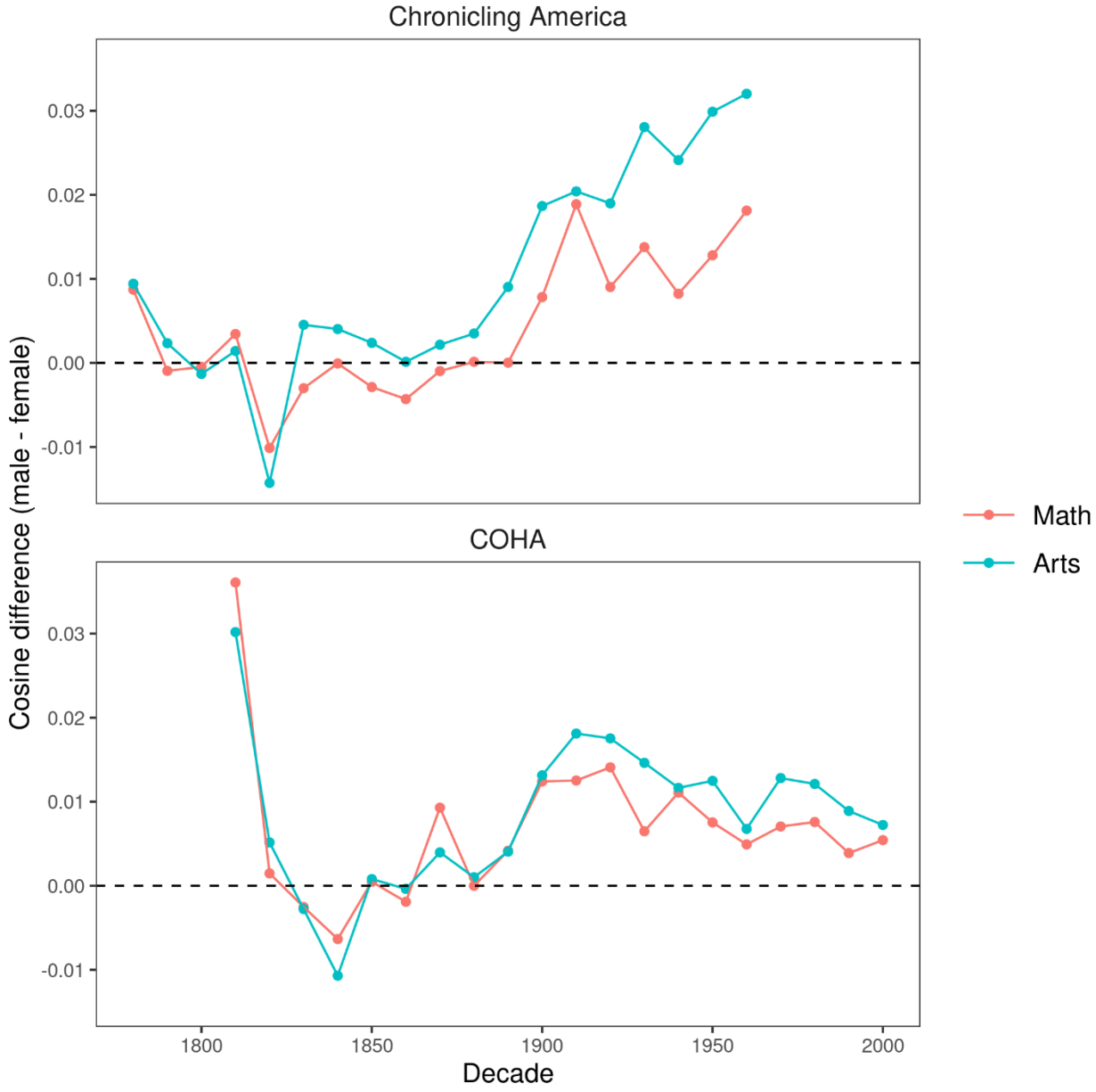


Figure 5.10 Results for experiment 7. The x axis shows decade and the y axis shows the difference in average cosine similarity between the male name concept vector and female name concept vector. The top subplot shows the results for the Chronicling America corpus and the bottom subplot shows the results for the COHA corpus. The points show the average cosine similarity difference color coded by each category of words (career, family).

In figure 5.10 we see a different pattern of results as the previous plots. For most decades across both corpora, the similarity between male names and math words is weaker than the similarity between female names and arts words. This relative ordering shows an unexpected pattern of hypothesized results. One reason for this could be that historically men had much greater representation as authors, poets, actors, and musicians causing male names to be more strongly related to both the math and arts words. The plots also show the male, career, math, and arts words become increasing related to the male names over time.

Across figures 5.8, 5.9, and 5.10 there are some interesting patterns of results. There is a sharp drop across all three plots for the Chronicling America corpus in the 1820s. One reason this could be is because of the increase discussion of women's rights around the 1820s.

Lastly, it is of note that there are pronounced differences between the patterns of results in the two corpora. One reason for the differences in patterns between the two corpora could be due to the types of documents found in the two corpora. The Chronicling America corpus features only American newspaper documents, whereas the COHA corpus features American newspapers, fiction, non-fiction, and magazines.

This experiment demonstrates a method of measuring theory driven research questions using semantic vectors over time. The method provides a way of measure changes in concepts or categories over time. Importantly, rather than comparing the similar between words, concepts are built up by summing (or averaging) the word vectors that are representative of a particular concept or category.

From a moral point of view these results are troubling. We know gender stereotypes have existed in the past and still occur in this day. Yet, we would hope that these stereotypes wouldn't be detectable in our writings, especially in the purported objective writings of newspapers that

aim to present facts. From a psychological view this analysis reveals that bias and stereotypes are embedded in the structure of our language. Even though the RP model is not human, has no brain, and is not belong to our social world, its simple associative mechanisms reproduced human stereotypes by merely being exposed to the structure of our language. An open question is whether these results are solely measuring a change in attitudes towards gender, or whether the results are also informed by *who* the writers are. Because the number of female writers in newspapers, fiction, and non-fiction has increased over time, these results might also be measuring changes not only in attitudes broadly but the source of the attitudes.

Experiment 8: Name Classification

The previous experiment shows great differences in male versus female names and what these names are related to, with male names being more strongly related to male relationships and pronouns, career, arts, and math, whereas female names are more strongly related to female relationships and pronouns, and family. The relations come from subtle patterns in the ways words are used together. To analyze differences in names more rigorously, I will use identical methodologies as in previous classification experiments to measure the extent that male versus female names are different and represent distinct category structures.

Names contain much more information than they would appear to at first glance, and indeed contain more information about the owner of the name than we would hope for in a fair and just society. For example, reading the names *Jamal*, *Emily*, or *Gladys*, likely recalls people of different races, genders, and generations. In this experiment, I wanted to measure differences between male and female names. This experiment can be thought of as a high dimensional t-test or ANOVA, where we can look for overall differences in 3000 numeric variables in several categorical groups. Importantly, I wanted to continue with the machine learning approach I have

used previous in this thesis, and focus on the predictive abilities of the model, rather than simply inferring predictive ability as is common with most tests of the null hypothesis (more on this in the general discussion).

Method. For this experiment, I used the top 200 male and female names (some of which appear as both male and female names) on the Social Security Services list of baby names from the 1880s to the 2000s.

For the first iteration of the experiment, I used the Chronicling America vectors that were derived from the 1880s. For each name, I extracted the corresponding word vector derived from the corpus. The result of this procedure is a 3000 dimensional space of 400 names where each name is represented by a point. Each name is labelled with either as male, female, or androgynous (names that appear in both the male and female name lists for a given decade). Figure 5.11 shows this 3000 dimensional space projected to 2 dimensions using Principal Components Analysis. Visually, the differences between the male and female names are apparent, with the two groups forming two overlapping clusters. Figure 5.12 shows the same space as figure 5.11, but zoomed in on the central cluster of names. Again, visually it is clear that these two groups of names are distinct to the model (androgynous names are not visually apparent).

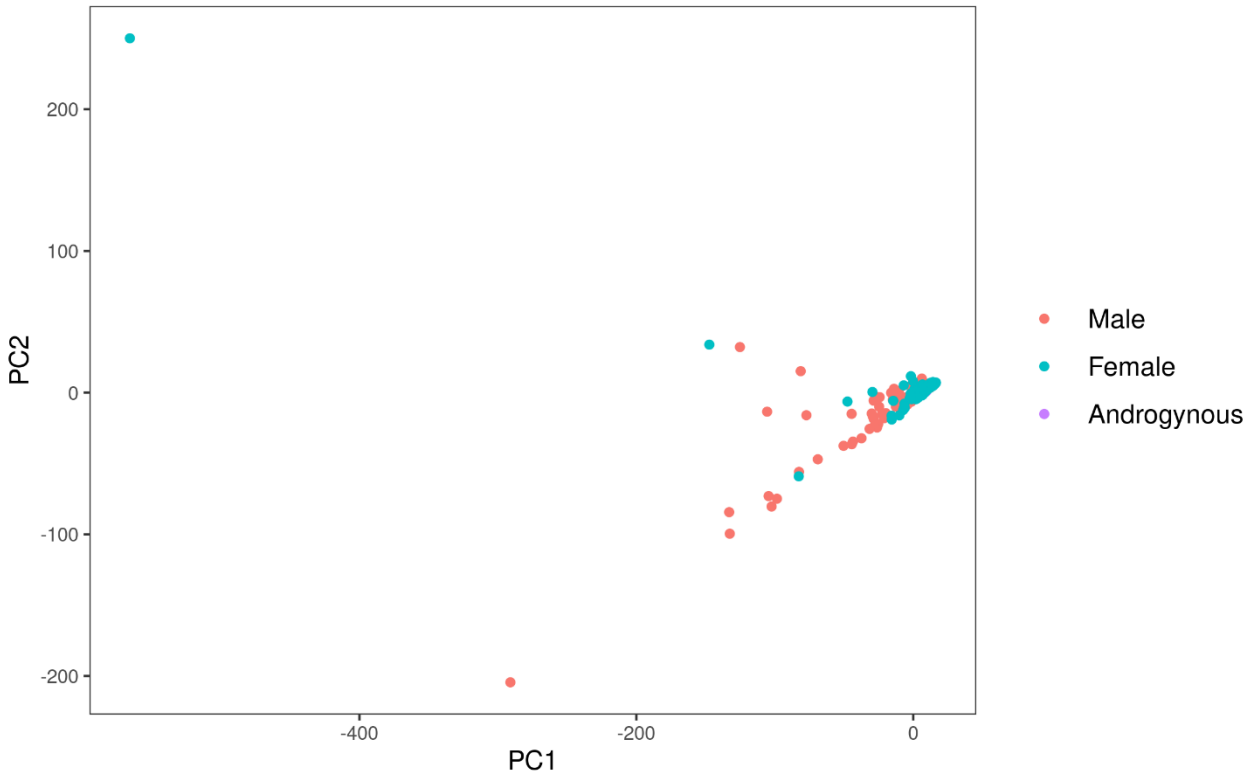


Figure 5.11 Two dimensional PCA rendering of 3000 dimensional space of male, female, and androgynous names from the Social Security Services list of most common baby names from 2000. The x and y axis show the first and second principal component of the 3000 dimensional space of names. Each point in the scatterplot represents one name color coded by gender (male: red; female: blue; androgynous: purple).

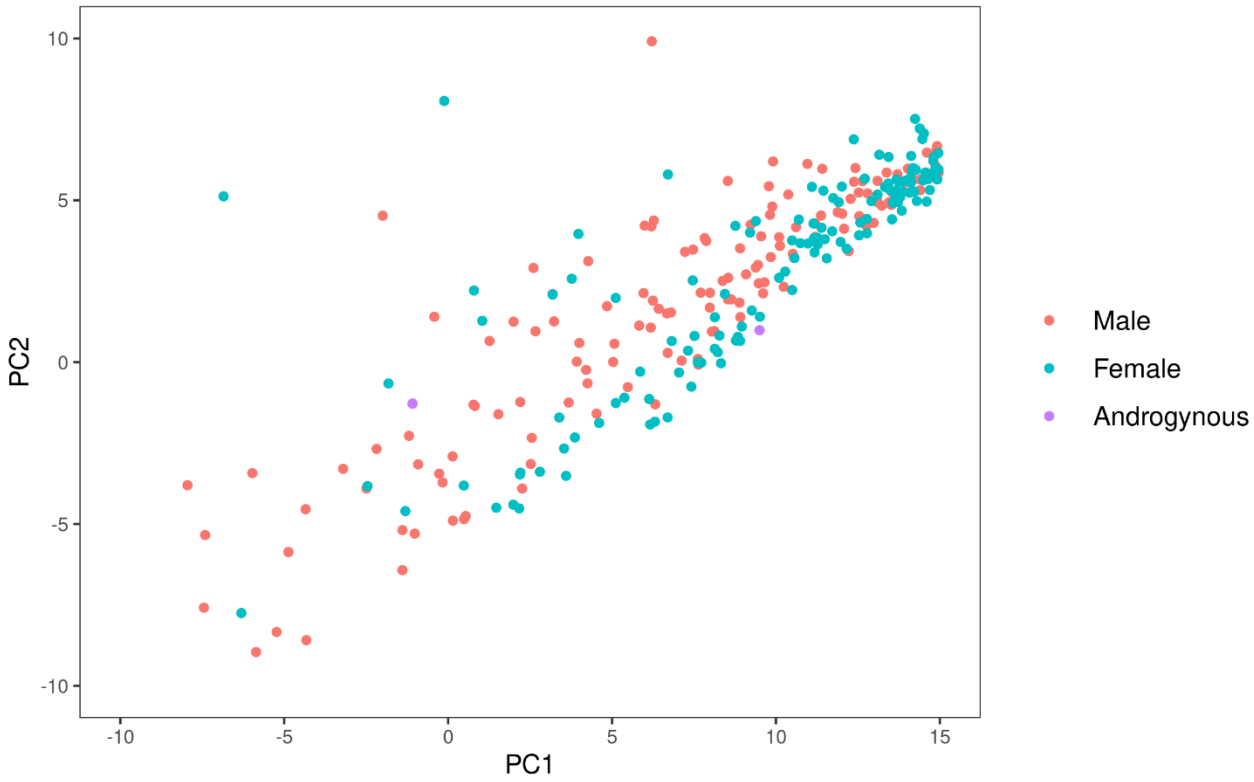


Figure 5.12 Two dimensional PCA rendering of 3000 dimensional space of male, female, and androgynous names. The figure shows the same data as figure 5.11, but zoomed in of the central cluster. The x and y axis show the first and second principal component of the 3000 dimensional space of names. Each point in the scatterplot represents one name color coded by gender (male: red; female: blue; androgynous: purple).

I wrote a program in R that split the dataset containing 400 names into 10 even groups (or folds to use the lingo of *k* fold cross validation). During each of $k = 10$ iterations, 360 labelled examples (i.e., the vector representation of the names with their associated category) were used to fit the Random Forest model. The model was then evaluated by being presented the remaining 40 unlabelled examples (i.e., the vector representation of the names without their associated

category) and the model made it classification decision for each of 40 unlabelled name vectors. The percent correct score was recorded.

The result of this procedure was 10 accuracy scores that were averaged to produce an unbiased estimate of the model's ability to classify novel names to one of the three predefined categories.

The program iterated though this procedure for all sets of decade vectors from the 1780s to 1960s for the Chronicling America corpus and all sets of decade vectors from the 1810s to 2000s for the COHA corpus. I evaluated the context vectors and the order vectors. The results for the order vectors for all experiments in this manuscript can be found in Appendix D.

Results. Figure 5.13 shows the decades from the 1880s to 2000s on the x axis and the model's average accuracy for a given decade on the y axis. Across every decade the model performed appreciably better than a chance model.

Discussion. The model was able to classify the gender of a name at a very high rate of accuracy across all decades of the corpora. The accuracy scores reporting in this experiment are higher than in even the simplest tests, such as the TOEFL or simple word classification tasks. These results show that there is a tremendous amount of structure in word vectors of names. The accuracy of the model is higher for the Chronicling America corpus. One reason this might be is that names are used in more stereotypical ways (i.e., there is a stronger and more distinct category structure in the newspapers in the Chronicling America corpus than in the COHA corpus that also contains fiction, non-fiction, and newspapers in addition to newspapers). This experiment demonstrates the extent to which gender is embedded in the words we use, such as our names.

From a psychological perspective these results reveal interesting information about the category structures of gender. After exposure to written language, the model sees male names as more similar to each other than compared to female names, and female names as more similar compared to male names. This demonstrates that our language is so well structured that it is almost too easy for this simple associative model to learn these well constructed social categories, just by being exposed to our written history.

Critically, in this experiment I measured the association between the vectors and their categories using out-of-sample data. I did not measure the relationship between the values of the vectors and their gender by standard statistical measures that have been shown to overfit the data. This is a point I will come back to in the general discussion.

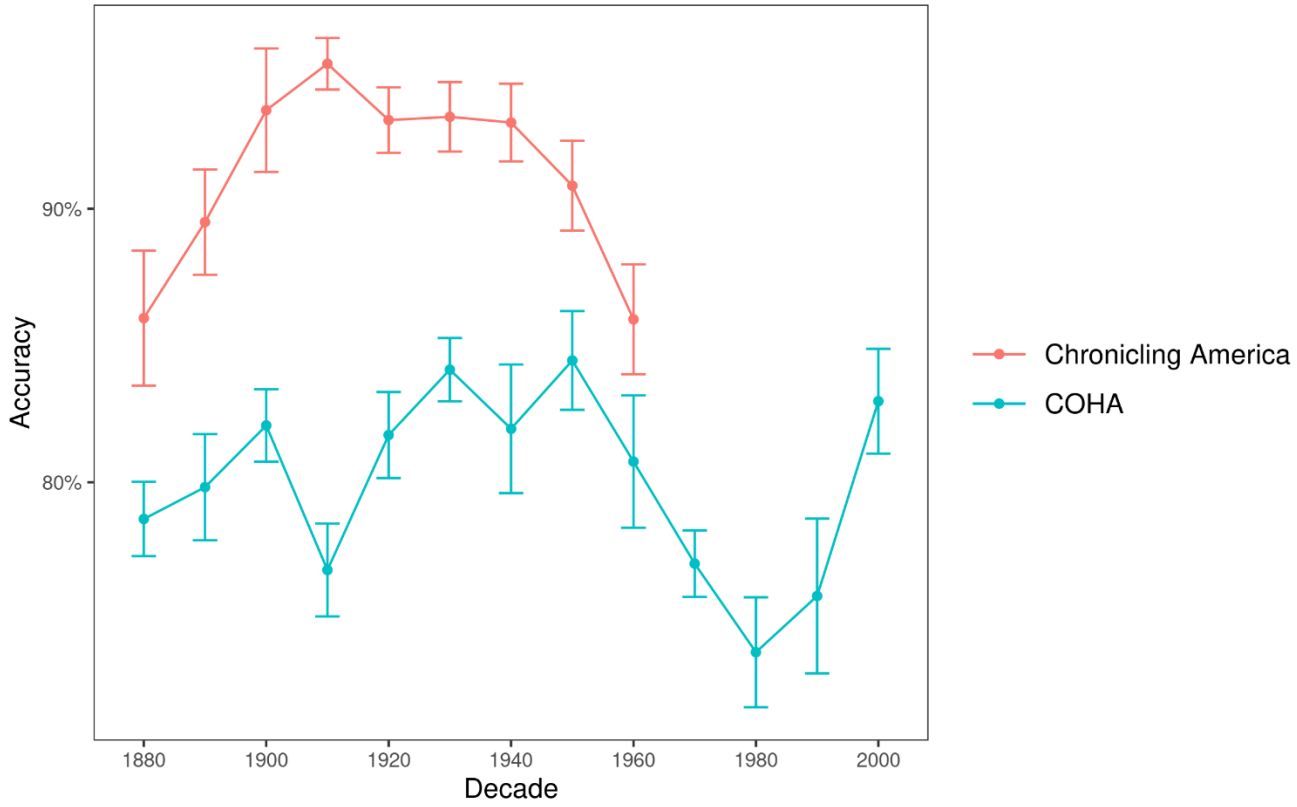


Figure 5.13 Results of experiment 8 across all decades. The x axis shows decade and the y axis is the accuracy of the model. Error bars show standard error of the mean.

General Discussion

The previous analyses came from several dozen sets of vectors from billions of words across over 200 years of American history using a combination of standard statistical methods (e.g., regression), semantic models (Random Permutation model), as well as machine learning models (e.g., Random Forests). I presented experiments ranging from gender bias to war, across 8 experiments and several dozen plots. However, there are a handful of key takeaways I want readers of this thesis to leave with.

First, as I argued at the beginning of this work, the language we use has a structure that reveals the structure of our thoughts. The structure shows the way in which our attitudes to historical events like war changes over time as well as our biases and stereotypical attitudes towards gender roles. This structure reveals itself in our writings as well as changes the way we think when we read the writing of others. All the methods I have used in this thesis have been used to detect subtle and not-so-subtle patterns in language. I programmed the Random Permutation model and developed vectors that encode the structure of language that makes language so informative and varied. The vectors contain – across their thousands of dimensions – the structure of human thought. Unfortunately, however, the meaning of the vectors is difficult to extract. To understand and extract the meaning of the vectors, as well as how they have changed over time, I used statistical learning models and developed methods to uncover the patterns and relationship across 3000 variables that represent the words we use to communicate everyday.

The second big takeaway of this project is on a more technical note. Very few papers have conducted text analysis on the Chronicling America database, and fewer still have used semantic models. The Chronicling America data is truly a massive dataset of raw and natural linguistic data (not pre-summarized counts). There are so many possible analyses that can be

conducted with the data, and I tried to show a few interesting ideas I had to present as case studies. But surely historians, economists, journalists, sociologists, and other psychologists will be able to think of more interesting, more specific, more impactful research questions to probe the large database for meaning in ways that are more impressive than imagined or can imagine. I have simply tried to advance the methods we have at our disposal to answer these questions. Researchers can use the vectors I have derived to answer cultural, historical, and societal questions using any programming language and methodologies.

Third, and as I also argued at the beginning of the thesis, the volume of data I have analyzed is more data than can ever be read by a single scholar or group of scholars. The subtle relationships and interactions that occur between the words and phrases across over 200 years of history cannot be understood by close reading. The only way to understand such a large volume of data is through the methods of statistics, modern data science, natural language processing, and machine learning. The statistical models like the Random Permutation model can understand complex relationships between words to build a representation of word meaning from direct cooccurrence (i.e., *dog* and *cat* often appear in the same sentence) as well as higher order cooccurrence (i.e., *dog* and *cat* each occur with another common set of words such as *pet* or *chase*). I argue that large scale semantic analysis is a useful and rigorous method of understanding the world we live in and how it has changed over time. Not only that, but automated methods of understanding data can be more accurate in understanding and predicting phenomenon as complex as the mental health and human behaviour (e.g., Meehl, 1954). Not only can these methods be more accurate, but we also tend to validate the methods in a more careful and rigorous way. For example, machine learning classification models are often evaluated for overall accuracy, as well as other measures of model quality such, as precision and

recall, or sensitivity and specificity. The models may also be tested to see how they classify within various social groups, such as how the accuracy of the model changes depending on which social group the data comes from. In contrast, human performance is typically judged much more crudely (i.e., humans perform one or several tests to qualify for licensure).

Furthermore, data science and statistics is really the best, if not the only, way to understand the world we live in and how it changes across time. It is the careful mining and analysis of data that helps us understand our history and our current place in the world. As an example, in the 1820s, 89% of the world's population lived in extreme poverty. In 2021, less than 10% of the world lives in extreme poverty. Depending on close readings of historical documents, our general sense about how poverty has changed, or news outlets to share this news of progress is simply not possible. It is only careful collection and interpretation of data, with the help of statistics, that allows us to understand and appreciate facts like this one out of dozens of examples of progress.

Understanding historical changes like the reduction of extreme poverty or our changing attitudes towards gender have benefits beyond simply being more aware of the state of the world we live in and how it has changed over time. After the realization that things used to be very different and inarguably worse in the not-so-distant past (e.g., almost 90% of the world lived in extreme poverty, women did not have the right to vote). The second realization we must confront is that in the short time following the not-so-distance past, *we* made the world very different and inarguably better over time. Billions of people were not lifted out of poverty or given the right to vote by accident. Rather, it was people that recognized problems in the workings of the world, whether those problems were problems we have been confronted with since the beginning of time, or problems we brought upon ourselves. The problems might have been lack of scarce

resources, soil that made growing food difficult, or attitudes and beliefs that were erroneous and misinformed. We then put forth hypotheses, either formally or informally on how we might change these problems. How can we find, produce, or use resources more effectively? How can we change attitudes toward racialized groups and improve intergroup relations more generally? Over our history we have come some distance is resolving these questions, or at least have caught a glimpse at what an answer might look like. We have used our knowledge to make the world better. And, given that we have made the world better in the past, there is hope we can continue to make the world better in the future. Our world has many problems, however, "Everything that is not forbidden by the laws of nature is achievable, given the right knowledge" (Deutsch, 2011). These problems and their solutions are interesting to study I their own right. But, the problems that are most interesting are also the most useful. Some of the most useful problems are those questions concerning how to reduce bias, eradicate discrimination, and improve social harmony. Statistics allows us to acquire knowledge to answer these questions about the world through careful and rigorous analysis.

This mention of statistics brings up a strange fact about this thesis – there are relatively few statistical tests reported given the quantitative and technical nature of this thesis. For example, I only reported *p* values a handful of times throughout the thesis and there is not one t test or ANOVA to be found. I deliberately avoided using statistical tests of the null hypothesis to the extent that I could throughout the entire thesis. It is important to remember that the purpose of conducting a test of the null hypothesis is to have confidence that the data collected are sufficiently ordered enough that if one was to collect more data of similar type that it would be ordered in approximately the same way. In other words, we come to believe the effect is real in

the sense that it is repeatable or reproducible. This is largely the logic behind fitting models in psychology and other disciplines.

However, when fitting models, Yarkoni & Westfall (2017) reminds us that "To fit is to overfit". Fitting a model means to extract structure from data in such a way that enables to us understand the structure of the data. If you understand the structure of data you can then make accurate predictions about future data (or out-of-sample data). To overfit is to extract structure from data that is too idiosyncratic to the data the model was fit to. These models do not allow for accurate predictions about future data (or out-of-sample data). To illustrate the point, Yarkoni & Westfall (2017) draws our attention to the titles of thousands of scientific papers that take the form "*x* predicts *y*…" such as "extraversion predicts performance in such and such memory experiment". And yet, the papers typically (aside from those choosing to use validation methods like I did in this thesis), do not predict anything in the standard use of the word predict. If one says they can predict the weather as an example, we rightly take that to mean that person can forecast weather that hasn't happened yet, not that they can explain why historical weather was a particular way. The distinction is not merely semantics. And yet, in psychology, we are typically in the explain-the-weather-after-the-fact scenario rather than demanding that our theories and models can actually predict data that our theories and models were constructed from.

Furthermore, even though this thesis is largely a theoretical and technical contribution, I believe there are many directly practical applications of this research outside of the realm of academics. For instance, the methods I used of measuring meaning and changes in meaning of male and female names could be used to measure bias and changes in bias in pronouns and names: In the case study I presented unbiased male and female vectors would be equally similar to the career, family, math, and arts concepts. This analysis could easily be conducted between

various newspapers or other publications to measure and understand bias. The analysis could also be used to measure differences in the way gender is presented between authors or within a single author's work across time. Similarly, in the name classification experiment I conducted, if there were no appreciable difference in how male and female names were used, the model would have a difficult time determining the difference between male and female names. This analysis could be put to practical use to analyze the meaning of specific male and female names, such as looking at how the meaning of names have changed overtime within a publication or how gender is represented in different occupations (e.g., nursing) or traits (e.g., agreeableness). In the political domain, Frimer et al. (2022), used similar computational methodologies to measure increases in political incivility of politicians through a text analysis of tweets. The RP model and related classification methods could be used for practical application to measure differences in meaning in political or religious language among different publications of political groups. These methods could allow got the careful measurement of shared and divergent meaning in contentious topics between different social groups.

Lastly, this thesis presents a potential cure to the issues of self report that psychology often relies on. For example, social psychologists rely on self report questionnaires to measure peoples' attitudes towards social groups (e.g., Shelton & Richeson, 2005). Yet, there are a host of known issues with using self reports, such as participants exhibiting biases like the social desirability bias, especially when the self report involves socially sensitive issues (e.g., Grimm, 2010). Methods that attempt to circumvent the problems of self report questionnaires, like the Implicit Association Test (IAT) proclaim to measure peoples' attitudes without their explicit permission (e.g., Craig & Richeson, 2014) also are not without their issues (e.g., Fiedler et al., 2006). The work in this thesis adds to a growing set of methodologies that can be used to

understand our attitudes, values, and beliefs, at both the levels of entire societies and cultures, or even at the level of individuals (if enough language data for a given subject is available). With more and more of our communication taking place online through text-based communication like email and text messages, in addition to public communication like social media (e.g., Twitter) becoming more common, these methods of extracting associations and meanings can be used to supplement or possibly even replace classic methods like self report questionnaires or the IAT.

Heuristics and Cognitively Inspired Approaches to Machine Learning

Cognitive psychology has long since established the importance of *pre*-dicting rather than *post*-dicting. Research on heuristic decision making (e.g., Gigerenzer & Brighton 2009) have evaluated the strategies that allow humans to understand the world so that they can successfully forecast the future accurately rather than strategies for making sense of the past. One of the big insights in this literature is that more data is not always better. For instance, making decisions using simple rules of thumb allows people to be more accurate than complex rules that try to model every aspect of the data. As an example, a sophisticated way of making a prediction (for example, whether it is going to rain tomorrow) is to consider all the relevant data available (whether it is raining now, the temperature, humidity, etc) and weight each predictor variable by its estimated importance. Indeed, this is the idea behind multiple regression and neural networks: model the dependent variable *y* as a function of a linear combination of independent variables ($b_0 + x_1b_1 + x_2b_2 + \ldots + x_nb_n$). However, a simpler approach is to forget the weighting of variable importance all together and simply tally up the variables to provide an estimate. Though the tallying method (or other simple heuristics such as a *choose the best* strategy) perform *worse on historical data* relative to multiple regression that weights each predictor by its measured importance, simple methods perform *better on novel data*.

To use the approach of measuring performance of a model or the structure in data by measuring the ability of the model to predict new data is very simple, straightforward, and unlike replicating an experiment, it doesn't depend on collecting any additional data. Validation methods all rely on resampling to fit the model to some portion of the entire dataset (e.g., 80%) and evaluate the fit on the remaining holdout portion (e.g., 20%). To give a concrete example, rather than conducting a simple linear regression and depending on a *p* value to evaluate whether the R^2 is significant and therefore trustworthy or real, one simply fits the regression model on 80% of the data and evaluates the R^2 by measuring the R^2 on the remaining 20%. Better yet, multiple resamples of the data can be generated, using a different 80-20% split each iteration. This produces a better estimate of the true R^2 or other effect size measurement. Many options for resampling methods are available such as bootstrapping, the jackknife, leave-one-out cross validation, and k fold cross validation. By design cross validation methods demand that a model perform well on novel data and one has to wonder if the discipline of psychology would have such concerns over replicability if we demanded that models actually predict data that they were not fit on.

Sarcasm, really?

A reasonable criticism of this work is that while I purport to be studying language using language models like the Random Permutation model, I have ignored some of the most interesting examples of language such as non-literal or figurative language (Sidtis & Sidtis, 2018). For example, the phrase *a dime a dozen* cannot be understood as a literal translation explaining the cost for twelve items. Figurative language is also more than the sum of its parts. The phrase like *beat around the bush* cannot be understood by analyzing the individual words but rather the phrase is understood holistically. Figurative language is not generated *on the fly* or

off the cuff, but rather is established through culture and committed to memory as a single chunk. Lastly, figurative language often carries emotional content. *It was a slap in the face* describes an action that someone perpetrates against another that is shocking and upsetting. Why did I not consider and model these interesting aspects of language in this work?

These features of figurative language make it some of the most memorable and poetic aspects of everyday language. However, given these features, figurative language is also very hard for language models to detect and even more difficult for them to understand in a meaningful and generalizable way. A couple of different approaches were available to me to model these examples of sarcasm, metaphor, irony, and idiom. First, hand crafted features can be generated to represent language that allow models to detect figurative language. Examples used in the machine learning literature are sentence length, hashtags, and exclamation points of a given text. For example, researchers have found sarcasm was reliably detected in a corpus of simple handcrafted features as I have described such as punctuation use. However, while machine learning models might be able to identify examples of figurative language, the models do not have a way to directly understand how this contributes or alters the meaning of the words in the phrase. These approaches also depend on a database of language examples that are labelled as formulaic versus non-formulaic expressions (e.g., sarcasm versus not sarcasm).

A more promising approach in the literature has been to use deep learning models to detect figurative language like metaphor (Harati et al., 2021). These models are neural network models with many hidden layers of artificial neurons that can model complex non-literal relationships. Unlike approaches described before that use hand crafted features, the models are fed raw text and learn from patterns in the text. However, like previously described methods, the model cannot be said to have any deeper understanding on what contributes to a phrase being

considered metaphor (for example). It is unclear how the classification of metaphor might contribute or alter the intended meaning of the passage. Furthermore, the deep learning approach still relies on a large, labelled database to train the model on. Given my thesis looks at changes in language over 200 years, to include measurements of formulaic examples of language I would have needed to compile a database of examples of formulaic language for each decade across nearly 200 years and train models for each decade to incorporate this information into the word and document representations.

I think it is a good reminder that whereas models like the Random Permutation model I used in this thesis, as well as semantic models more generally (like LSA, BEAGLE, GLOVE, etc), are sometimes called *language models,* I believe that it is an error to call these models language models because semantic models are not truly models of language. Language has many aspects that contribute to meaning, such as semantics, syntax, punctuation, misuses of punctuation to communicate alternative meanings (i.e., emojis like :)), and more. Language also depends on cultural and historical background knowledge to understand language in a more holistic way. Understanding non-literal forms of language is possible due to this deeper level of knowledge. Sarcasm, irony, humour, idioms, and more are a critical aspect of language, not only for meaning, but are shared within a culture and also make language a form of art and individual expression.

It is important to remember that semantic models like the Random Permutation model are not models of language, but rather are models of *an aspect of language*, namely semantics. Granted, semantics is an important aspect of language, but the model misses out on some truly important aspects of language such as figures of speech, irony, and sarcasm that depend on historical events, societal factors, and culture.

Biased People, Biased Data, Biased Models

Humans are biased in their thinking and speech. This leads our writing to be biased. We have stereotypical assumptions about who is a doctor, nurse, mechanic, receptionist, or soldier, and we reinforce those biases again in our writing which influences others. A great deal of research has gone into catalogue, understand, and even correct bias in text corpora. A reasonable question that can be asked of this thesis is why I didn't try to correct for some known biases in the dataset I am using, or build vectors that have less noticeable bias. In my project, I had very few text pre-processing methods other than selecting from a standard pre-defined list of words. But, if as the saying goes, garbage in, garbage out, why didn't I try to clean up the garbage in my data so that I might lead to vectors that have less exhibitable bias?

There is a great deal of work that has gone into methods of debiasing text corpora. For example, these methods can balance text corpora so that words representing particular social groups (e.g., men and women) more equally co-occur with various concepts such as career and family. Furthermore, many useful methods exist for debiasing vectors so that they keep their useful semantic properties (e.g., the vector for male is more similar to the vector for father than mother) while ridding the model of its negative properties (e.g., the vector for male being more similar to the vector for doctor than nurse).

The corpus I used in this thesis and the vectors I have derived from it are biased. The fact that decade after decade men and women are stereotypically associated with drastically different social roles demonstrates measurable bias that cannot be ignored. Critically however, this bias is not due to arbitrary biases like more newspaper articles coming from a particular time period, geographic region, or due to particular text processing choices I made. Rather, the bias I have been able to measure is systematic bias that reflects the bias we exhibit in our written language.

This bias is the exact type of bias that opens up a large database of text to psychology to understand human emotion, thought, and behaviour.

Future directions

Hyperparameter tuning and comparing semantic models. Many papers using machine learning models (semantic models or otherwise) often focus on placing various models in competition with each other on a range of tasks. The models are compared to determine which one might be "best" or "state of the art". In this thesis, I have shied away from making model comparison the bulk of the thesis. Rather, I decided to choose an established, psychological grounded model to develop methods of measuring meaning across time. Similarly, many papers spend an inordinate amount of time tuning modifiable hyperparameters. Unlike regular parameters in statistic models (e.g., the slope and intercept in a regression models), hyperparameters cannot be directly estimated from the data, but rather have to be evaluated on some measurement of error or fit. For example, the random permutation model has several hyperparameters: vector length, number of non-zero elements, window size, and the composition of the environment vectors. Though testing a very large number and selecting the best performing set of hyperparameters might have given me better results in some of my experiments, finding an optimal set for a given task or set of tasks was never my intention with this thesis. For one, as I have previously mentioned, it is possible (and likely) to overfit data and it is unlikely the time and pages I would have spent explaining and tuning parameters would have contributed anything meaningful to the development and testing of the methods I have outlined in this thesis. For two, I wanted to use a fairly standard set of hyperparameters to focus on the application of measuring meaning over time, and not call into question whether the results of the entire thesis were because of a carefully selected set of hyperparameters for the semantic

model I used. Similarly, though I have no way to prove this in writing, I did not try dozens of different experiments that did not work as intended and exclude them from this thesis, only showing the handful of useful things that "worked out". This so-called *file drawer problem*, where scientists are said to have a file drawer full of failed experiments for every one experiment that is a success, is a problem across all science, but I think its an especially pernicious potential problem in computational work where ideas can be tried out so quickly and easily. I did my best to avoid this, therefore I did not try dozen of models and experiments.

That being said, I think future work should investigate many different models from psychology, such as LSA, BEAGLE, HAL, and LIWC across this massive corpus of historical text. These models might very well perform better than the random permutation model I used. Similarly, supervised models common outside of psychology like word2vec, GLOVE, BERT, and others may also be very useful for measuring meaning over time. Fitting these models across a range of parameters might produce results that are far greater than I have produced in this thesis, and might even be stable across a wide range of tasks for measuring meaning throughout the decades of the corpus.

Making the vectors and analyses available to the research community. The questions I set out to answer in this thesis and the number of analyses I conducted is only a microscopic fraction of all possible questions and analyses possible given such a large and rich database of text. Given that I cannot ask every conceivable question from such a rich source of data, I am going to make the vectors I have derived available to data scientists, computer scientists, historians, sociologists, and psychologists who can then ask any research question of interest. Once downloaded, researchers can analyze the vectors using any standard computer language such as Java, R, Python, or any other language or software they choose.

However, because not all researchers have the programming experience required to work with the raw vectors, I am also going to build a web-based interface similar to Google Books N-Gram viewer. The interface will allow researchers to investigate any research question they have using the same methods I describe in this thesis (e.g., measuring valence or word strength association) using an intuitive point and click interface to generate a graph of results as Graphical User Interfaces (GUIs) are becoming increasing popular methods of allowing researchers to analyze semantic word data (e.g., Lutfallah et al., 2018).

Conclusion

This project used two large corpora of newspaper (and other text) data spanning over 200 years of American history. These newspapers chronicle many major events in American history: the rise and fall of presidents and political parties, two world wars, the abolishment of slavery, the rights movement of women, children, and non-human animals, and inventions that allow our species to put a human on the moon and cook food using invisible waves. The newspapers also describe events that may never make history books, but are just as interesting: a runaway slave, a woman on her seventh hunger strike in support of the suffragette movement, and the promise of a $50 reward for any information on the whereabouts of a stolen horse. Importantly, this record of events, both large and small, can give us a glimpse into what life was like in the past. More interestingly, we can use computational tools and theory from psychology to excavate these archeological records of language and psychology to understand the stereotypes, associations, concepts, and ideas of the people of the past. Data science, statistics, and computational modelling allows us to compare to these findings of the past to ourselves, not only giving new understanding of people from the past, but also of ourselves in the present.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In Proceedings of Human Language Technologies: *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19-27). Association for Computational Linguistics.

Aujla, H. (2021). Language experience predicts semantic priming of lexical decision. Canadian *Journal of Experimental Psychology*, *75*(3), 235–244. https://doi.org/10.1037/cep0000255

Aujla, H., Jamieson, R. K., & Cook, M. T. (2018). A psychologically inspired search engine. *Lecture notes in computer science: high performance computing systems and applications*.

Aujla, H., Crump, M. J. C., Cook, M. T., & Jamieson, R. K. (2019). The Semantic Librarian: A search engine built from vector-space models of semantics. *Behavior Research Methods*, *51*, 2405-2418.

Baker, S., Reichart, R., & Korhonen, A. (2014). An unsupervised model for instance level subcategorization acquisition. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 278-289).

Balogh, E. P., Miller, B. T., & Ball, J. R. (2015). *Improving diagnosis in health care.* The National Academies Press.

Bedi, G., Carrillo, F., Cecchi, G., Slezak, D., Sigman, M., Mota, N., . . . Corcoran, C. (2015). *Automated analysis of free speech predicts psychosis onset in high-risk youths*. *NPJ Schizophrenia*, *1*(1), 150-130.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, *94*(4), 991-1013.

Bradley, M.M., & Lang, P.J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical report C-1, Gainesville, FL. *The Center for Research in Psychophysiology*, University of Florida.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional semantics in technicolor. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*(pp. 136-145). Association for Computational Linguistics.

Buchanan, L., Westbury, C. & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. Psychonomic Bulletin & Review, 8, 531–544. https://doi.org/10.3758/BF03196189

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet 7*(2):223–242.

Caliskan, A., Bryson, J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* (New York, N.Y.), *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Craig, M. A., & Richeson, J. A. (2014). More diverse yet less tolerant? How the increasingly diverse racial landscape affects white Americans' racial attitudes. *Personality and Social Psychology Bulletin*, *40*(6), 750-761.

Darwin, C. (1902). *The descent of man*. New York: American Home Library.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. *In Proceedings of the international AAAI conference on web and social media*. *11*(1), 512-515).

Denhière, G., Lemaire, B., Bellissens, C., & Jhean, S. (2008). A semantic space for modeling children's semantic memory. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of latent semantic analysis*. 143-165. Mahwah, NJ: Laurence Erlbaum and Associates.

Deutsch, D. (2011). *The beginning of infinity: Explanations that transform the world*. Penguin UK.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *WWW*, 29–30.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

Eichstaedt, J., Smith, R., Merchant, R., Ungar, L., Crutchley, P., Preoţiuc-Pietro, D., … Eichstaedt, J. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(44), 11203–11208. https://doi.org/10.1073/pnas.1802331115

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The journal of machine learning research*, 15(1), 3133-3181.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European review of social psychology, 17*(1), 74-147.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes, 25*(2-3), 285-307.

Frimer, J. A., Aujla, H., Feinberg, M., Skitka, L. J., Aquino, K., Eichstaedt, J. C., & Willer, R. (2022). Incivility Is Rising Among American Politicians on Twitter. Social Psychological and Personality Science. https://doi.org/10.1177/19485506221083811

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science, 1*(1), 107-143.

Greenfield, P. (2013). The Changing Psychology of Culture From 1800 Through 2000. *Psychological Science*, *24*(9), 1722–1731. https://doi.org/10.1177/0956797613479387

Grimm, P. (2010). Social Desirability Bias. In J. Sheth, & N. Malhotra (Eds.), *Wiley International Encyclopedia of Marketing*. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781444316568.wiem02057

Gustafsson Sendén, M., Lindholm, T., & Sikström, S. (2014). Selection Bias in Choice of Words: Evaluations of "I" and "We" Differ Between Contexts, but "They" Are Always Worse. *Journal of Language and Social Psychology, 33*(1), 49–67. https://doi.org/10.1177/0261927X13495856

Halawi, G., Dror, G., Gabrilovich, E., & Koren, Y. (2012). Large-scale learning of word relatedness with constraints. *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1406-1414). ACM.

Harati, P., Westbury, C., & Kiaee, M. (2021). Evaluating the predication model of metaphor comprehension: Using word2vec to model best/worst quality judgments of 622 novel metaphors. *Behavior Research Methods, 53*(5), 2214–2225. https://doi.org/10.3758/s13428-021-01558-w

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, *23*(6), 1744-1756.

Holtgraves, T. M. (2013). *Language as social action: Social psychology and language use.* Psychology Press.

Howard, M. W., Addis, K. A., Jing, B., & Kahana, M. J. (2007). Semantic structure and episodic memory. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 121–141). Mahwah, NJ: Laurence Erlbaum and Associates.

Hill, F., Reichart, R., & Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *4*, 665-695.

Jacobs, J. B., and Potter, K. 2000. *Hate crimes: Criminal Law and Identity Politics*. Oxford University Press

Johns, B., & Dye, M. (2019). Gender bias at scale: Evidence from the usage of personal names. *Behavior Research Methods, 51*(4), 1601–1618. https://doi.org/10.3758/s13428-019-01234-0

Johns, B. T., & Jamieson, R. K. (2019). The influence of time and place on lexical behavior: A distributional analysis. *Behavior Research Methods*, *51*, 2483-2453.

Johns, B. T., Taler, V., Pisoni, D. B., Farlow, M. R., Hake, A. M., Kareken, D. A., ... & Jones, M. N. (2018). Cognitive modeling as an interface between brain and behavior:

Measuring the semantic decline in mild cognitive impairment. *Canadian Journal of Experimental Psychology, 72*(2), 117.

Jones, M. N., & Mewhort, D. (2007). Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychological Review*, *114*(1), 1-37.

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer & J. T. Townsend (Eds.). *Oxford Handbook of Mathematical and Computational Psychology*, 232-254.

Jurafsky, D., & Martin, J. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, N.J: Pearson Prentice Hall.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty : heuristics and biases*. Cambridge, UK: Cambridge University Press.

Khoo, C., Chan, S., & Niu, Y. (2002). The many facets of the cause-effect relation. T*he Semantics of Relationships* (pp. 51-70). Springer, Dordrecht.

Kuhn, M., & Johnson, K. (2016). *Applied Predictive Modeling*. New York: Springer.

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

Kwantes, P. Derbentseva, N. Lam, Q. Vartanian, O. & Marmurek, H. (2016). Assessing the Big Five personality traits with latent semantic analysis. *Personality and Individual Differences*, *102*, 229-233.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *AAAI*.

Landauer, Thomas K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*(2), 211-40.

Landauer, T. K., Laham, R. D. & Foltz, P. W. (2003). Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. In M. Shermis & J. Bernstein, (Eds.), *Automated Essay Scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Landsdall-Welfare , T., Sudhahar, S., Thompson, J., Lewis, J., Team, F. N., & Cristianini, N. (2017). Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences*, *114*(4), E457-E465.

Lindquist, K. A., MacCormack, J. K., & Shablack, H. (2015). The role of language in emotion: Predictions from psychological constructionism. *Frontiers in psychology*, *6*, 444.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

Lutfallah, S., Fast, C., Rangan, C., & Buchanan, L. (2018). Semantic neighbourhoods: There's an app for that. The Mental Lexicon.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction*, *21*, 251–284.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T., Xu, F., & Clahsen, H. (1992). Overregularization in Language Acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), i–178. https://doi.org/10.2307/1166115

Martin, D. I., & Berry, M. W. (2011). Mathematical Foundations Behind Latent Semantic Analysis. In T. K. Landuaer, D. S. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*, (pp. 35-55). Mahwah, NJ: Laurence Erlbaum and Associates.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN, US: University of Minnesota Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Millis, K., Magliano, J., Wiemer-Hastings, K. Todaro, S., & McNamara, D. S. (2007). Assessing and Improving Comprehension with Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 121–141). Mahwah, NJ: Laurence Erlbaum and Associates.

Moretti, F. (2013). *Distant reading*. London: Verso.

Newman, M., Pennebaker, J., Berry, D., & Richards, J. (2003). Lying words: Predicting deception from linguistic styles. *Personality And Social Psychology Bulletin*, *29*(5), 665–675. https://doi.org/10.1177/0146167203251529

Nosek, B. A., Banaji, M., and Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration web site. Group Dynamics: Theory, Research, and Practice, 6(1):101.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Review*, *49*, 197–237.

Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pincombe, B. (2004). Comparison of human and latent semantic analysis (LSA) judgements of pairwise document similarities for a news corpus. *Defence Science And Technology Organisation Salisbury (Australia) Info Sciences Lab.*

Pinker, S. (2003). Language as an adaptation to the cognitive niche. *Studies in the Evolution of Language*, 3, 16-37.

Pinker, S. (2013). Learnability and Cognition, new edition: *The Acquisition of Argument Structure.* MIT press.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*, 623–641.

Recchia, G., Sahlgren, M., Kanerva, P., & Jones, M. (2015). Encoding Sequential Information in Semantic Space Models: Comparing Holographic Reduced Representation and Random Permutation. *Computational Intelligence and Neuroscience*, (2015), *18*. https://doi.org/10.1155/2015/986574

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*, 627-633.

Sahlgren, M. (2005). An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE)*, Copenhagen, Denmark.

Sahlgren, M., Host, A., & Kanerva, P. (2008). Permutations as a means to encode order in word space. In V.Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 1300–1305). Mahwah, NJ: Lawrence Erlbaum Associates.

Shannon, C., & Weaver, W. (1949). The mathematical theory of communication. Urbana: University of Illinois Press.

Shelton, J. N., & Richeson, J. A. (2005). Intergroup contact and pluralistic ignorance. *Journal of personality and social psychology*, *88*(1), 91.

Silge, J. & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *Journal of Open Source Software*, 1(3), 37–. https://doi.org/10.21105/joss.00037

Sidtis, D. V. L., & Sidtis, J. J. (2018). Cortical-subcortical production of formulaic language: A review of linguistic, brain disorder, and functional imaging studies leading to a production model. *Brain and cognition*, *126*, 53-64.

Strang, G. (1998) *Introduction to Linear Algebra*. Wellesley, MA: Wellesley Cambridge Press.

Tausczik, Y., & Pennebaker, J. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., … Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, *571*(7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8

Turney, P., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research, 37*, 141-188.

Vapnik V. (2010). *The Nature of Statistical Learning Theory*. New York, NY: Springer New York.

Warriner, A., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

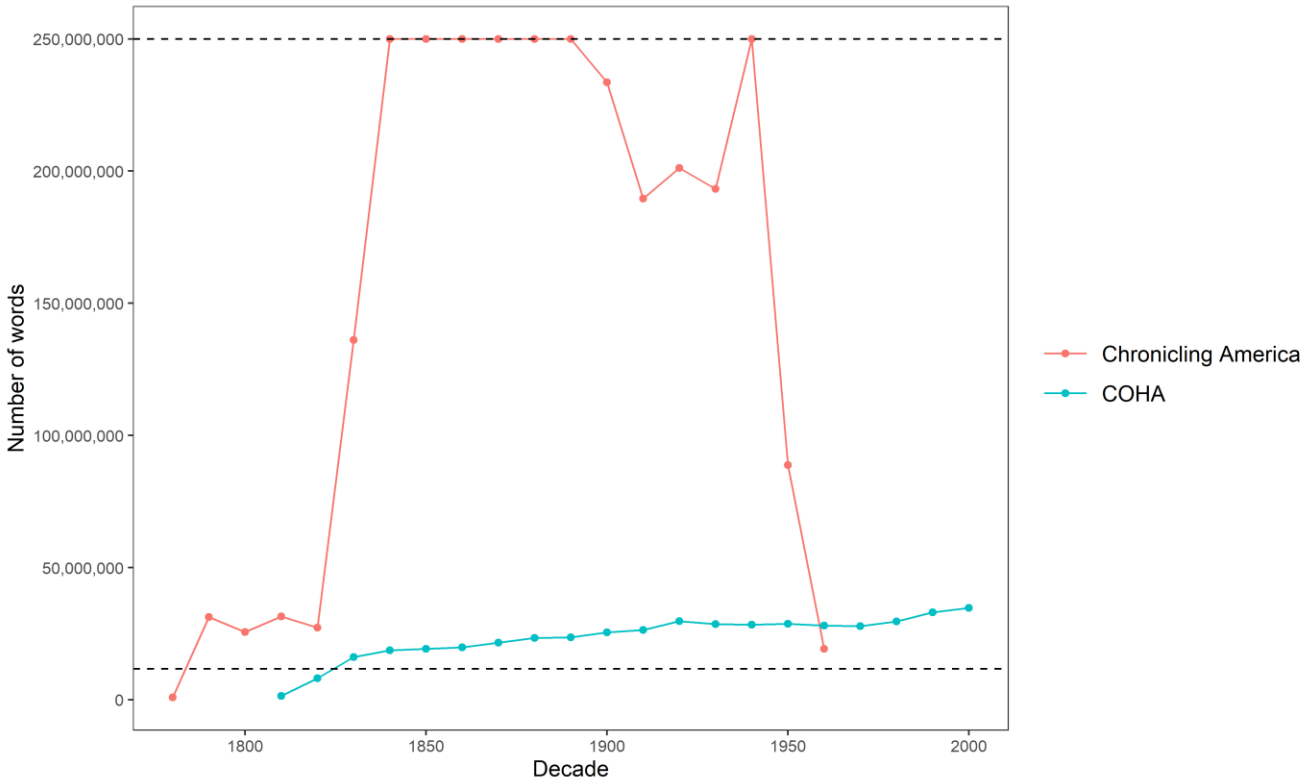
Widdows, D. (2004). *Geometry and meaning*. Stanford, Calif.: CSLI Publications, Center for the Study of Language and Information.

Willits, J. A., Rubin, T., Jones, M.N., Minor, K. S., & Lysaker, P. H. (2018). Evidence of disturbances of deep levels of semantic cohesion within personal narratives in schizophrenia. *Schizophrenia Research*, *197*, 365-369.

Yarkoni, T. (2010). Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44*(3), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100-1122.

Appendix A



Word counts for Chronicling America and COHA corpora. The horizontal line at 250 million words indicates that decades with more than 250,000,000 words were limited to 250 million words. The horizontal line at approximately 11 million represents the number of words in the TASA corpus for comparison.

Appendix B

Materials for experiment 3

countries: china, india, indonesia, pakistan, brazil, nigeria, bangladesh, russia, mexico, japan

animals: dog, cat, wolf, elephant, fish, snake, mouse, bear, fox, deer

body parts: eyes, ears, nose, mouth, hand, arm, leg, shoulder, finger, toe

colors: red, green, yellow, blue, black, white, purple, pink, brown, grey

sports: golf, baseball, football, hockey, badminton, basketball, volleyball, tennis, soccer, rugby

vegetables: carrot, potato, tomato, corn, onion, lettuce, pepper, broccoli, celery, mushroom

fruits: banana, apple, strawberry, grape, pear, watermelon, blueberry, lemon, peach, cherry

diseases: chickenpox, influenza, malaria, measles, meningitis, mumps, polio, virus, cancer, pneumonia

professions: doctor, mechanic, nurse, driver, painter, clerk, assistant, supervisor, bookkeeper, carpenter

furniture: table, chair, couch, desk, ottoman, stool, bookcase, seat, bed, cabinet

Appendix C

Materials for experiment 5

Each participant saw each of the ten descriptions below twice in random order, once with the male name and pronoun and once with the female name and pronoun.

Noah/Emma is a recent MBA graduate from the University of Manitoba. **He/She** takes a hands-on approach to managing employees and describes **him/her** self as outgoing and likeable. **His/Her** hobbies include going out on the weekend with friends and traveling.

Liam/Sophia is about to graduate from the University of Winnipeg. **His/Her** says motivation is the most important aspect of managing teams. **His/Her** hobbies include reading and trying new restaurants around Winnipeg.

Jacob/Olivia graduated with a business degree from the University of Brandon. **He/She** constantly works to motivate employees through positive feedback and incentives. **His/Her** spends **his/her** free time reading and studying leadership strategies.

Mason/Isabella is a recent business graduate from the University of Manitoba. **He/She** takes a hands-off approach to leadership and describes **Him/Her** self as tough but fair to employees. **His/Her** interests include reading and watching tv.

William/Ava graduated with a PhD in psychology from University of Manitoba. **He/She** has a hands-on approach to running a business. **His/Her** hobbies are public speaking and writing.

Ethan/Mia has no post secondary education but is enthusiastic and energetic. **Ethan/Mia** says he/she always works to motivate employees to perform their best. **He/She** plays the piano and enjoys watching movies.

Michael/Abigail graduated with a business degree from the University of Manitoba. **He/She** focuses on communication and understanding while managing employees. **His/Her** hobbies include rock climbing and mountain biking.

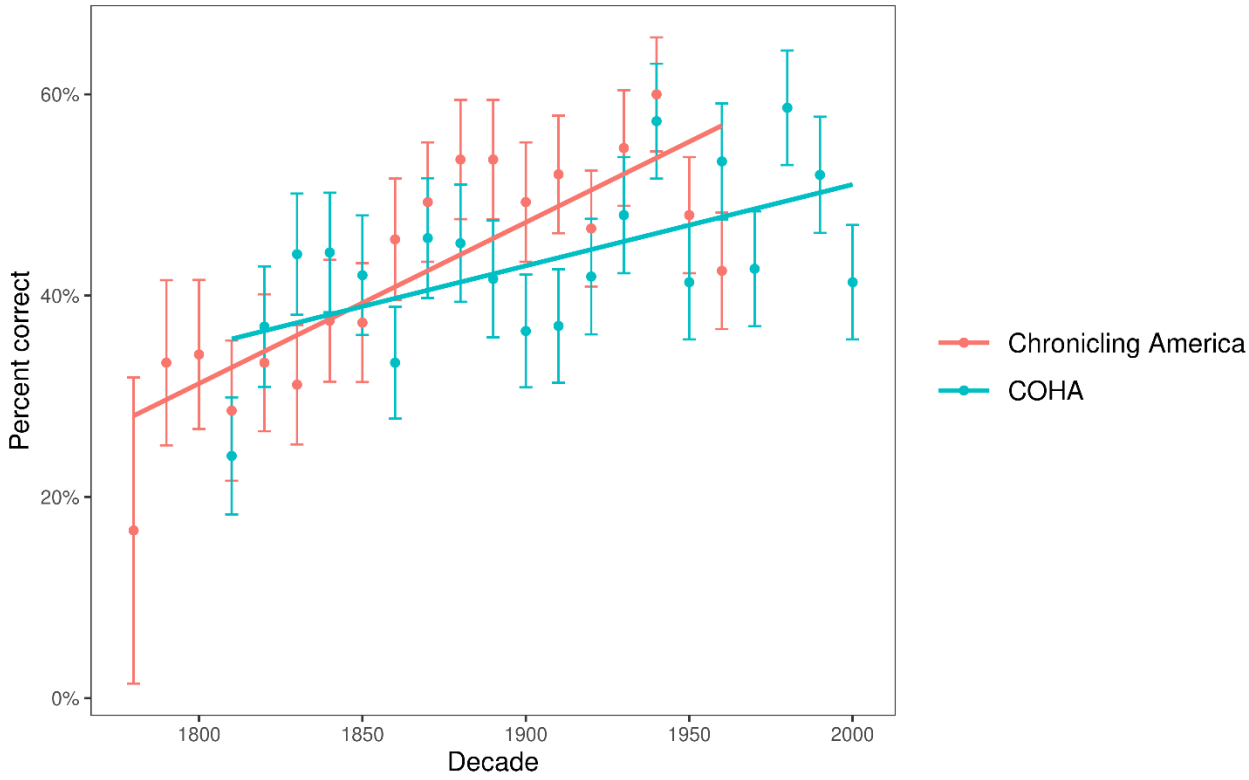
Alexander/Emily graduated with a master's degree in psychology from University of Manitoba. **He/She** main focuses is building a business on strong values. **His/Her** hobbies are public speaking and volunteering for local charities.

James/Madison graduated with a business degree from the University of Winnipeg. **He/She** approach to leadership is to carefully track employee outcomes. **His/Her** interests are reading about business strategy.

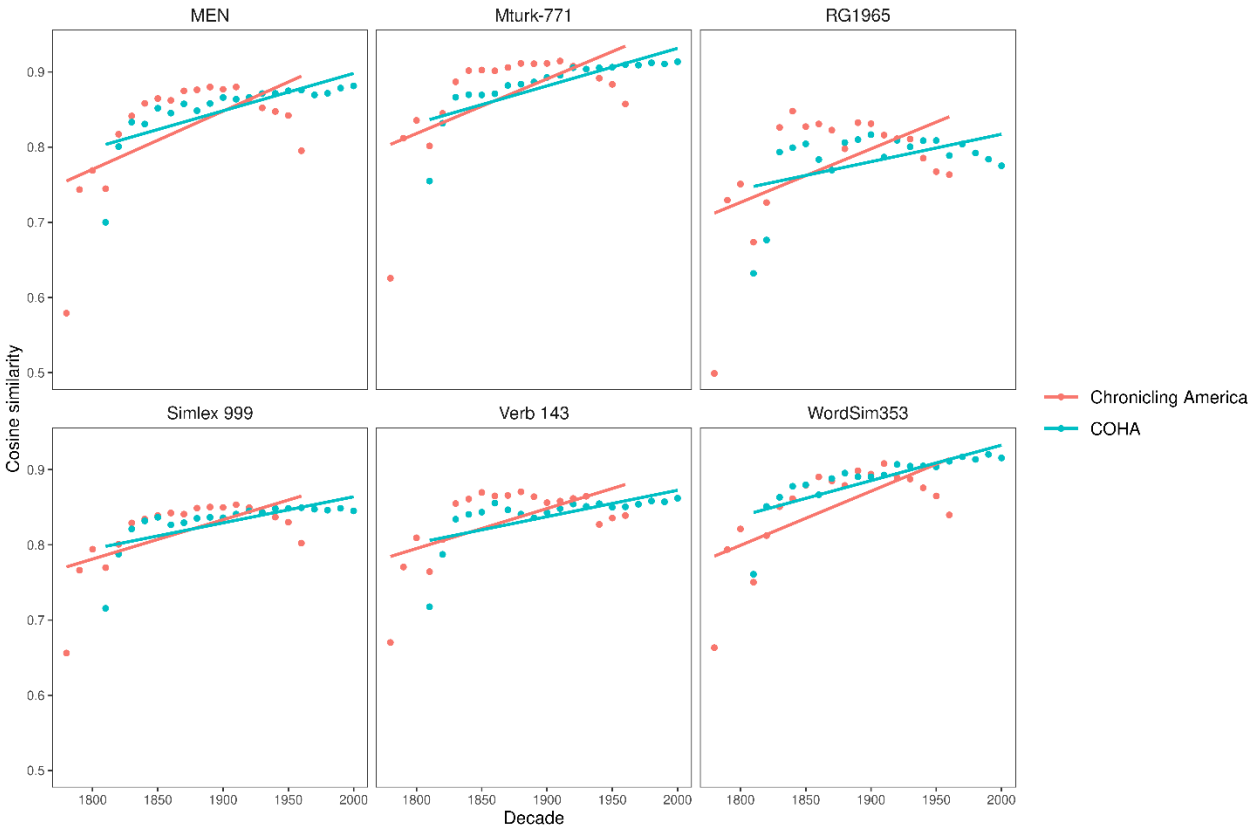
Elijah/Charlotte graduated with a PhD in psychology from University of Manitoba. **He/She** focuses on bring out the best in employees while managing a business. **His/Her** hobbies include camping and painting.

Appendix D

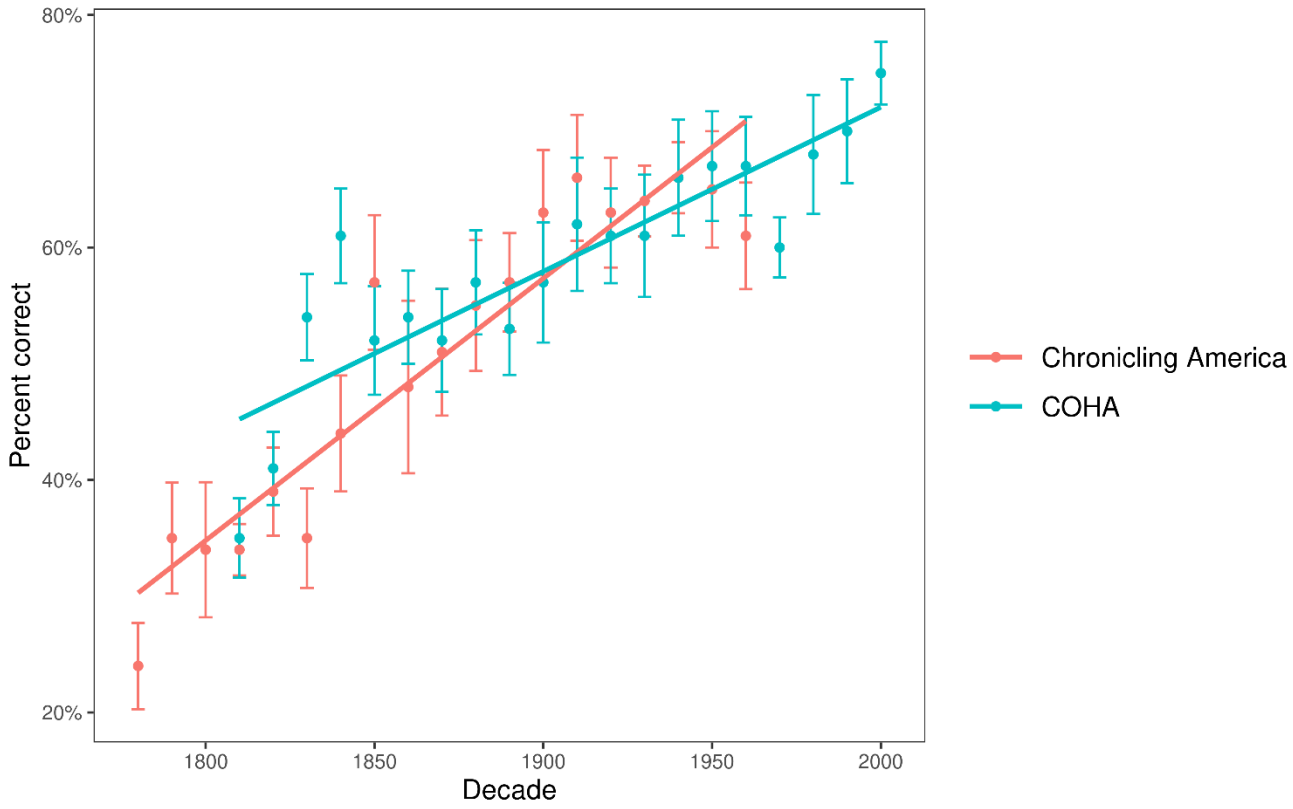
Replication of all experiments using order vectors



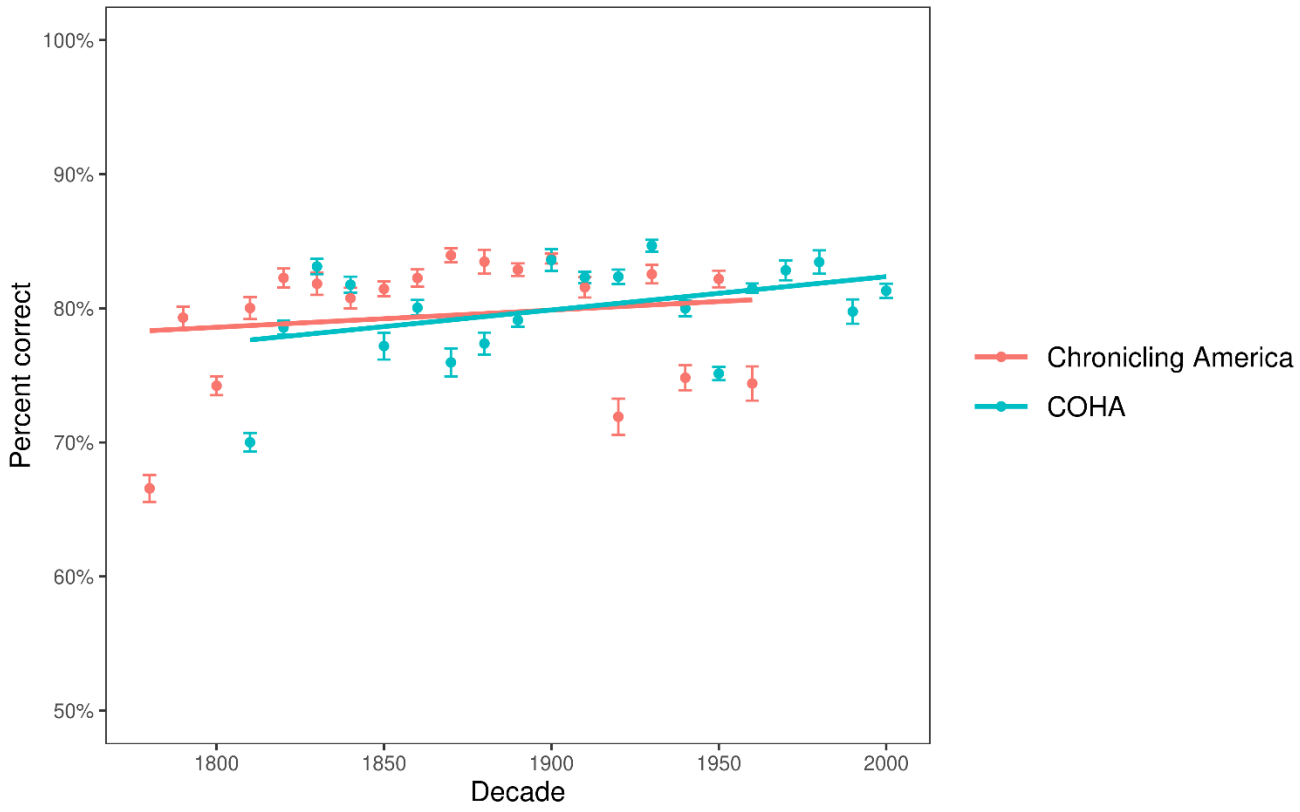
Results from Experiment 1. The x axis displays the decade and the y axis displays the percent correct. Each point represents the accuracy for a given decade on the TOEFL and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade. The error bars show the standard error of the proportion.



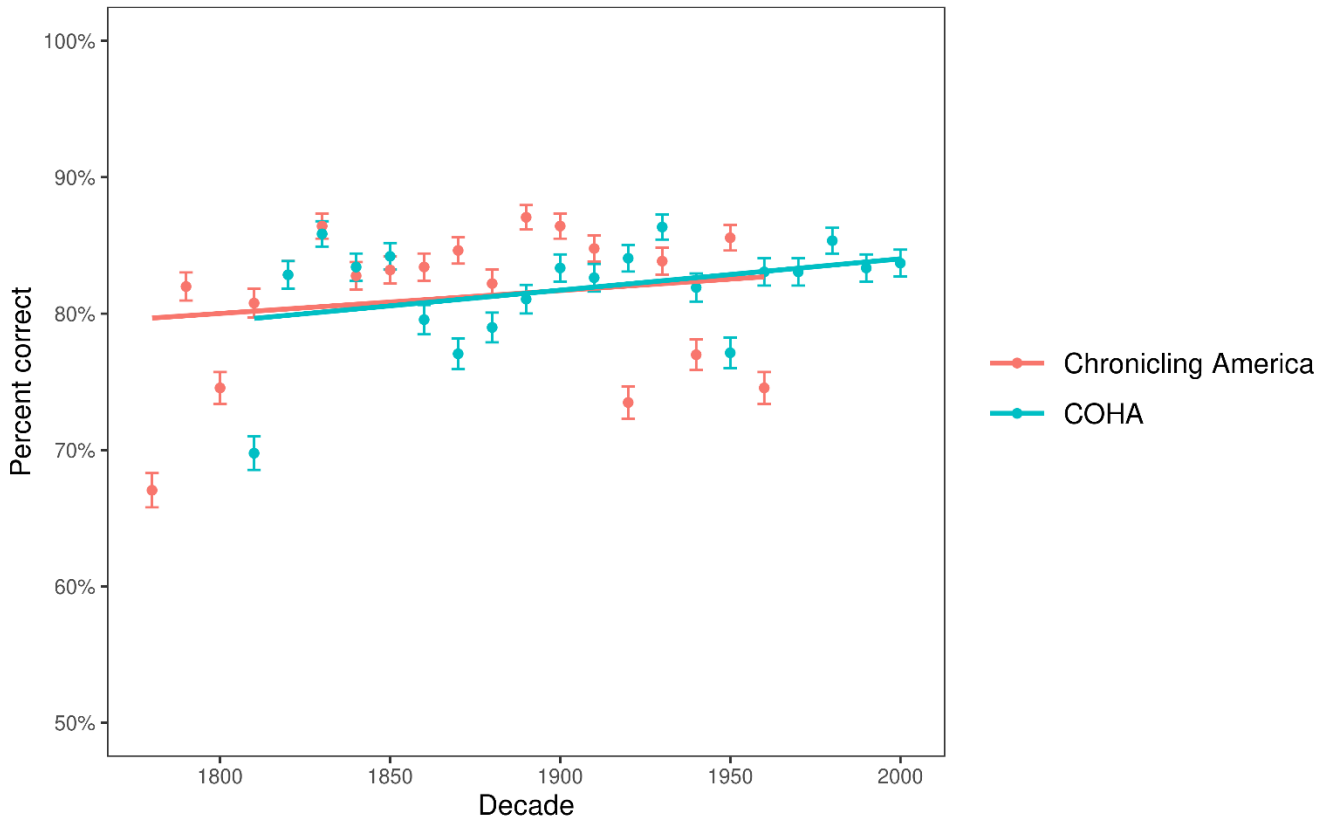
Results from experiment 2. In each of the six subplots, the x axis displays the decade and the y axis displays the cosine similarity for six different word similarity test. Each point represents the cosine similarity for a given test for a given decade and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade.



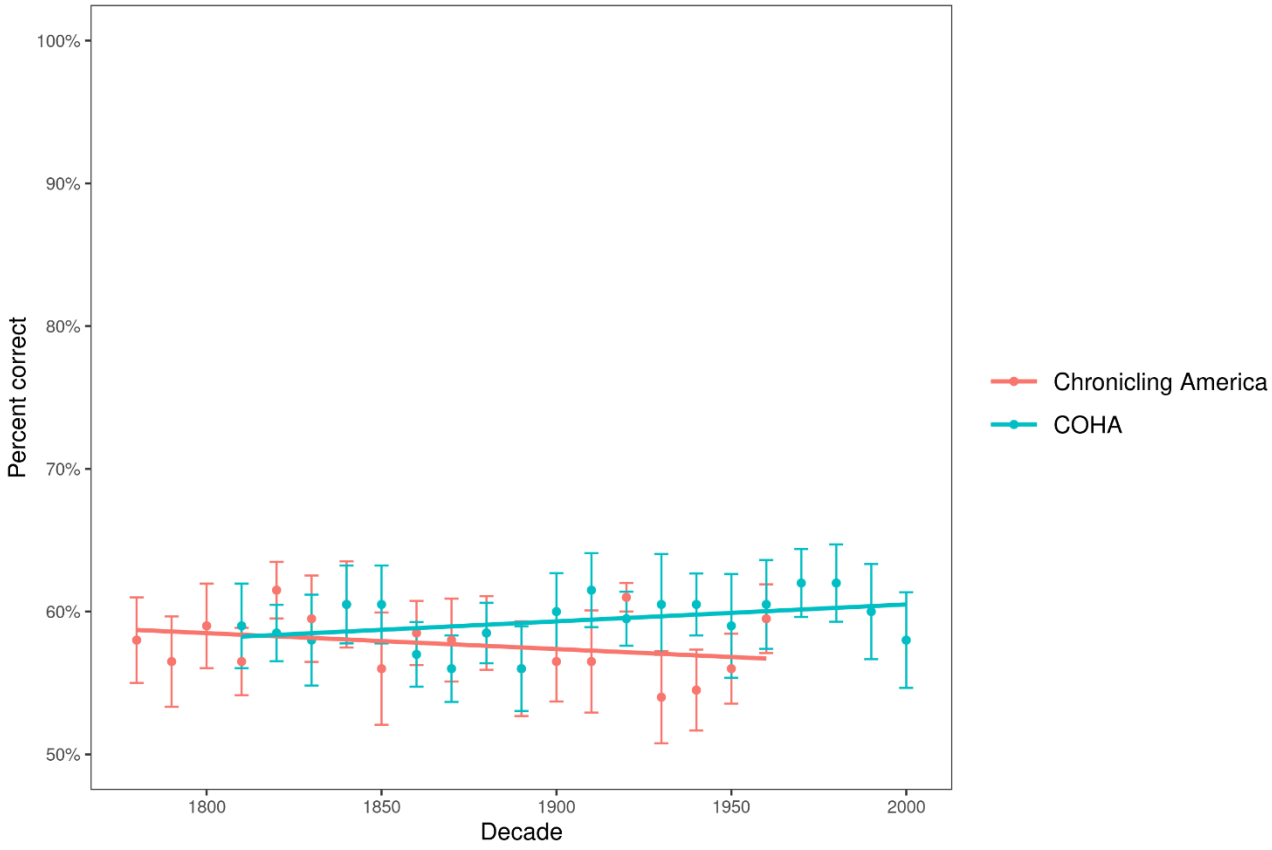
Results from experiment 3. The x axis displays the decade from which the vectors were derived and the y axis displays the percent correct for the word classification task. Each point represents the accuracy for a given decade and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade. The error bars show standard error of the mean.



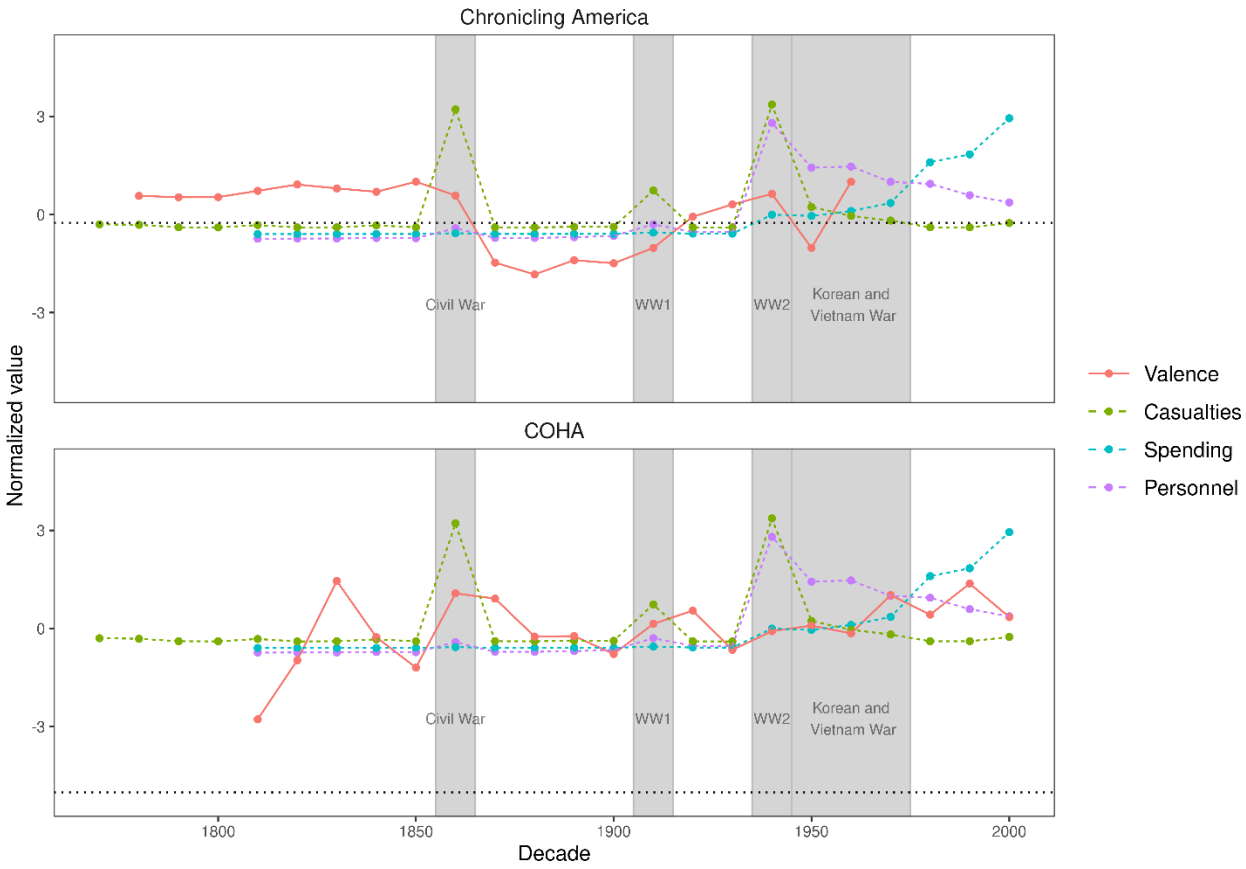
Results from experiment 4 (using the cross validated training data). The x axis displays the decade and the y axis displays the percent correct for the Twitter hate speech classification task. Each point represents the accuracy for a given decade and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade. Error bars show the standard error of the mean.



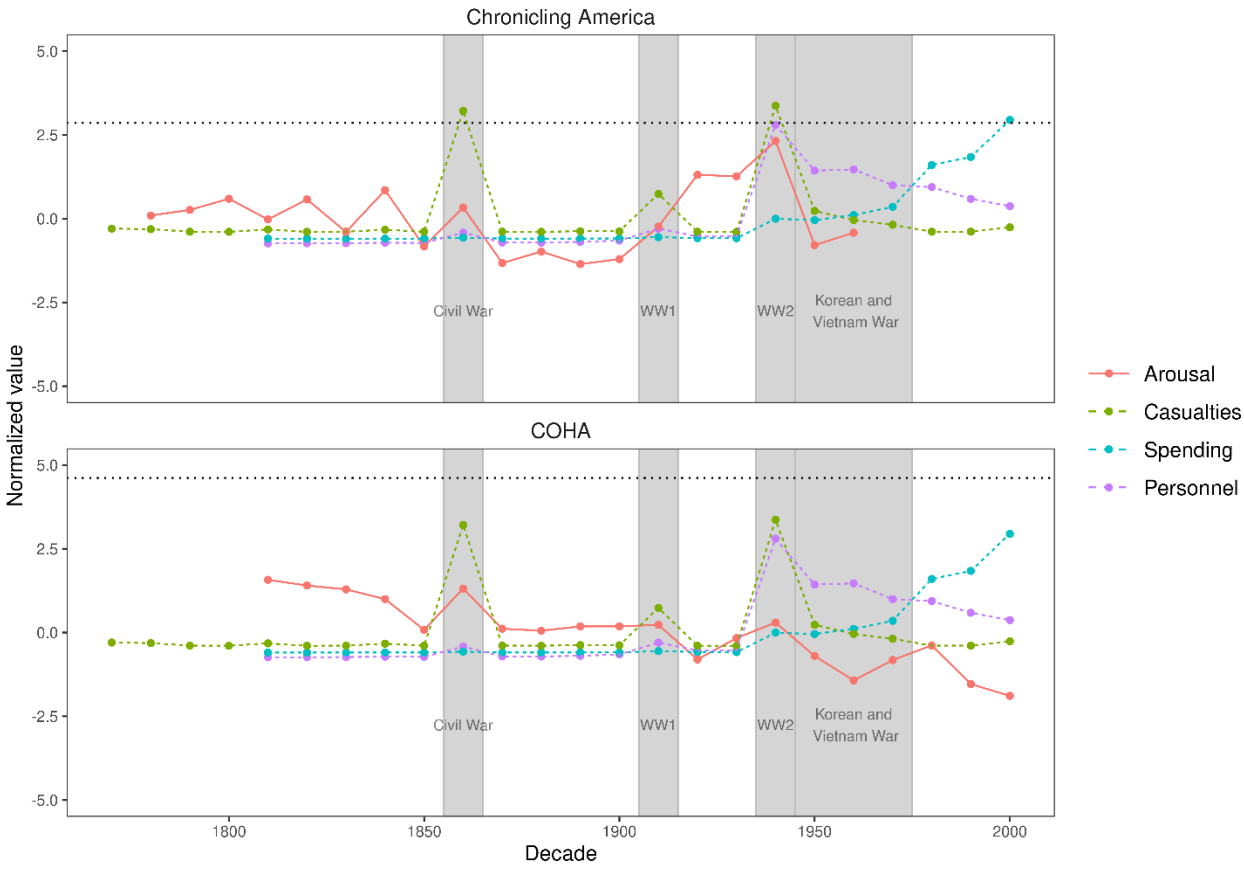
Results from experiment 4 (using the test data). The x axis displays the decade and the y axis displays the percent correct for the Twitter hate speech classification task. Each point represents the accuracy for a given decade and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade. Critically, the accuracy reported in this plot is computed using data the model was not trained on. The error bars show the standard error of a proportion.



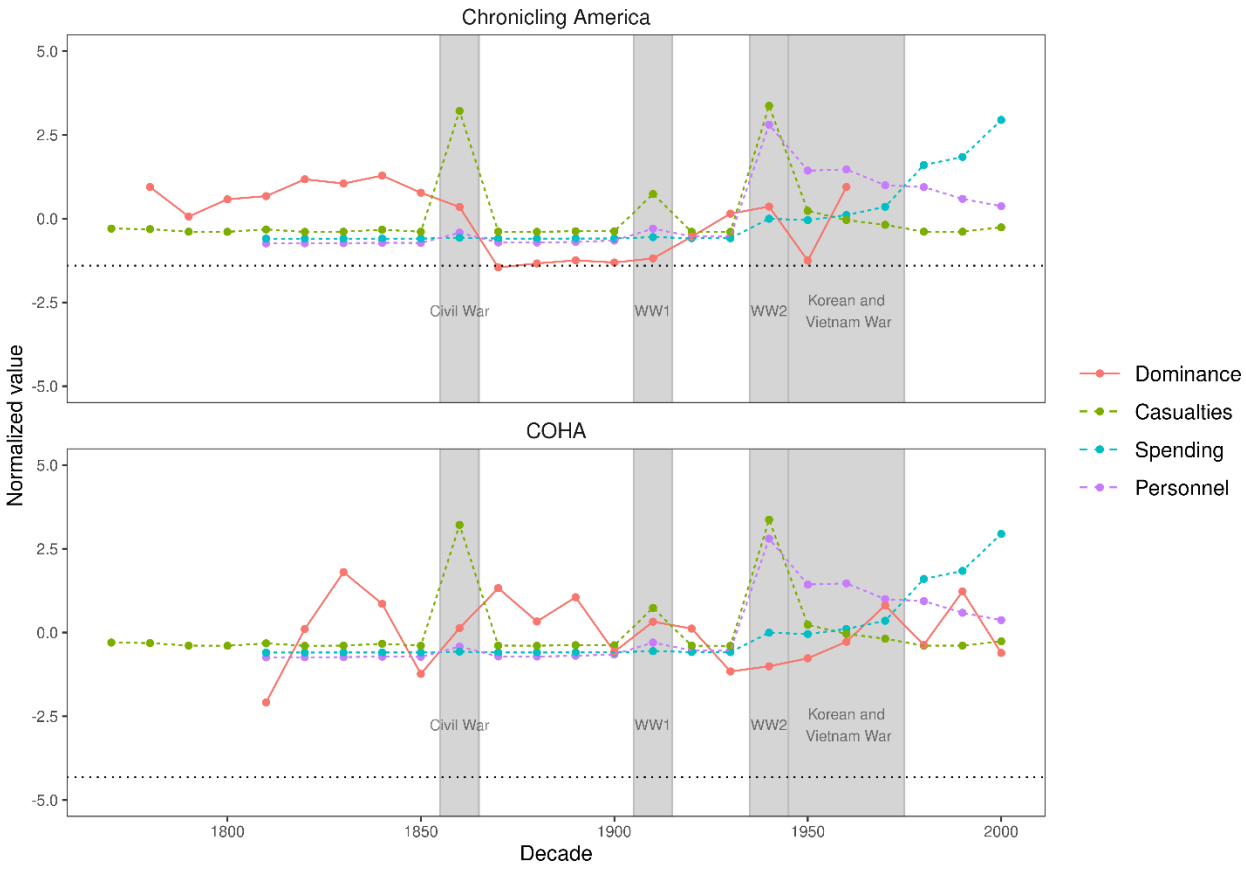
Results from experiment 4. The x axis displays the decade and the y axis displays the percent correct for the hiring bias classification task. Each point represents the accuracy for a given decade and are color coded by corpus (where red = Chronicling America and blue = COHA). The two straight lines show the least squares regression line for percent correct as a function of decade. Error bars show the standard error of the mean.



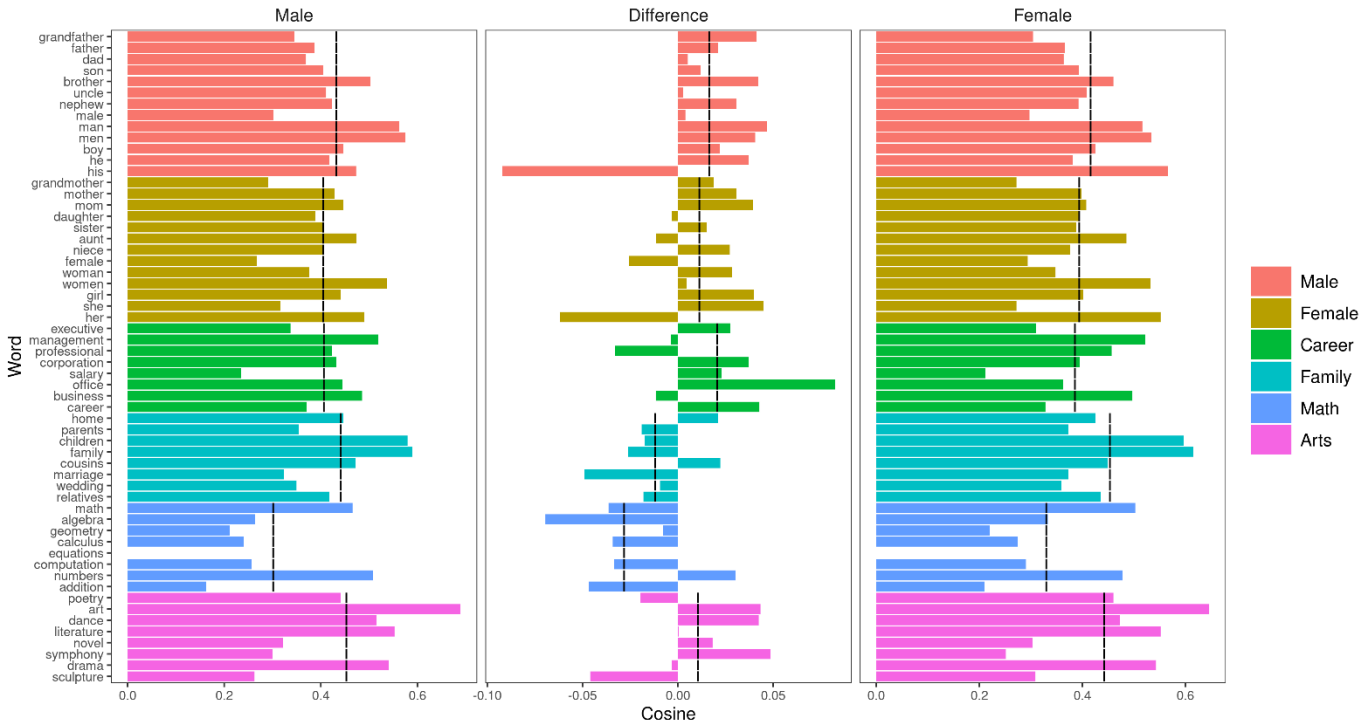
Results from experiment 6 for valence. In the two subplots, the x axis displays the decade and the y axis displays the normalized measurement of valence and normalized empirical measurements (war casualties, spending, and personnel). The top subplot displays the results for the Chronicling America corpus and the bottom subplot displays the results for the COHA corpus. All variables are standardized to offer easy visual comparison (i.e., all four variables are converted to z scores). The horizontal dotted line represents the z score corresponding to the center of the valence scale (e.g., a 5 on the 9-point Likert-type scale).



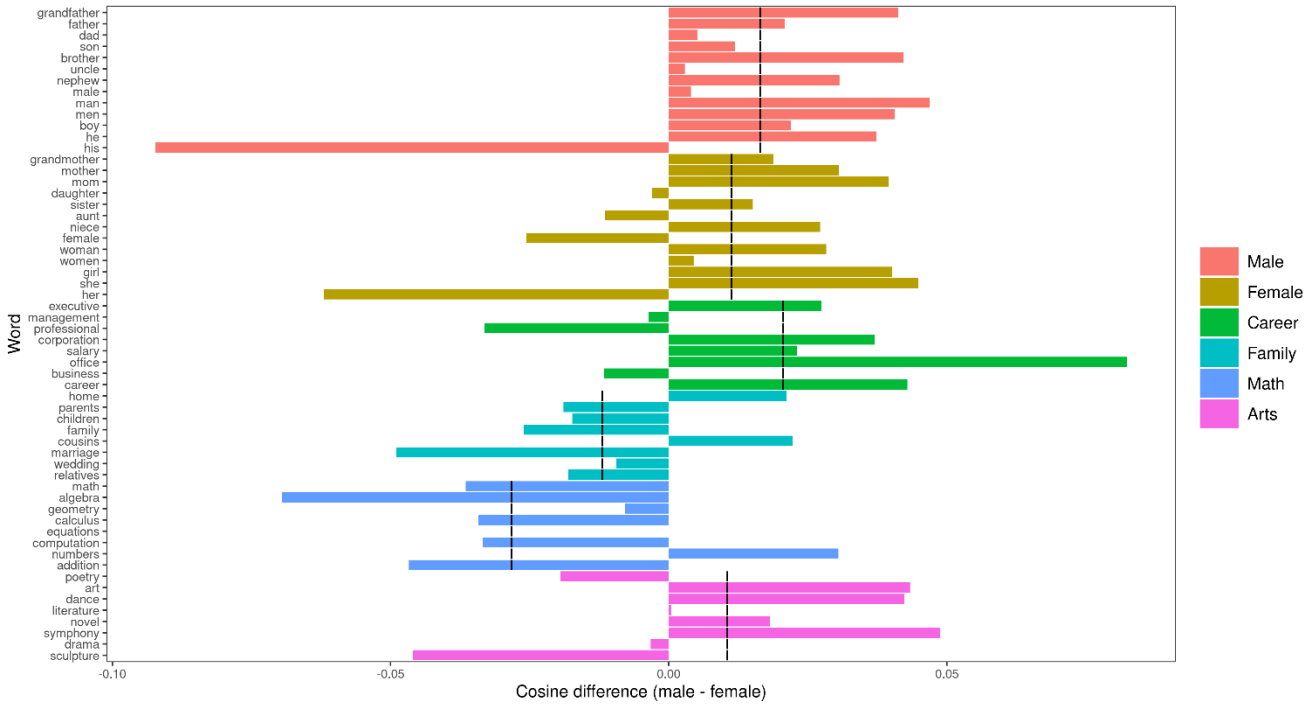
Results from experiment 6 for arousal. In the two subplots, the x axis displays the decade and the y axis displays the normalized measurement of arousal and normalized empirical measurements (war casualties, spending, and personnel). The top subplot displays the results for the Chronicling America corpus and the bottom subplot displays the results for the COHA corpus. All variables are standardized to offer easy visual comparison (i.e., all four variables are converted to z scores). The horizontal dotted line represents the z score corresponding to the center of the arousal scale (e.g., a 5 on the 9-point Likert-type scale).



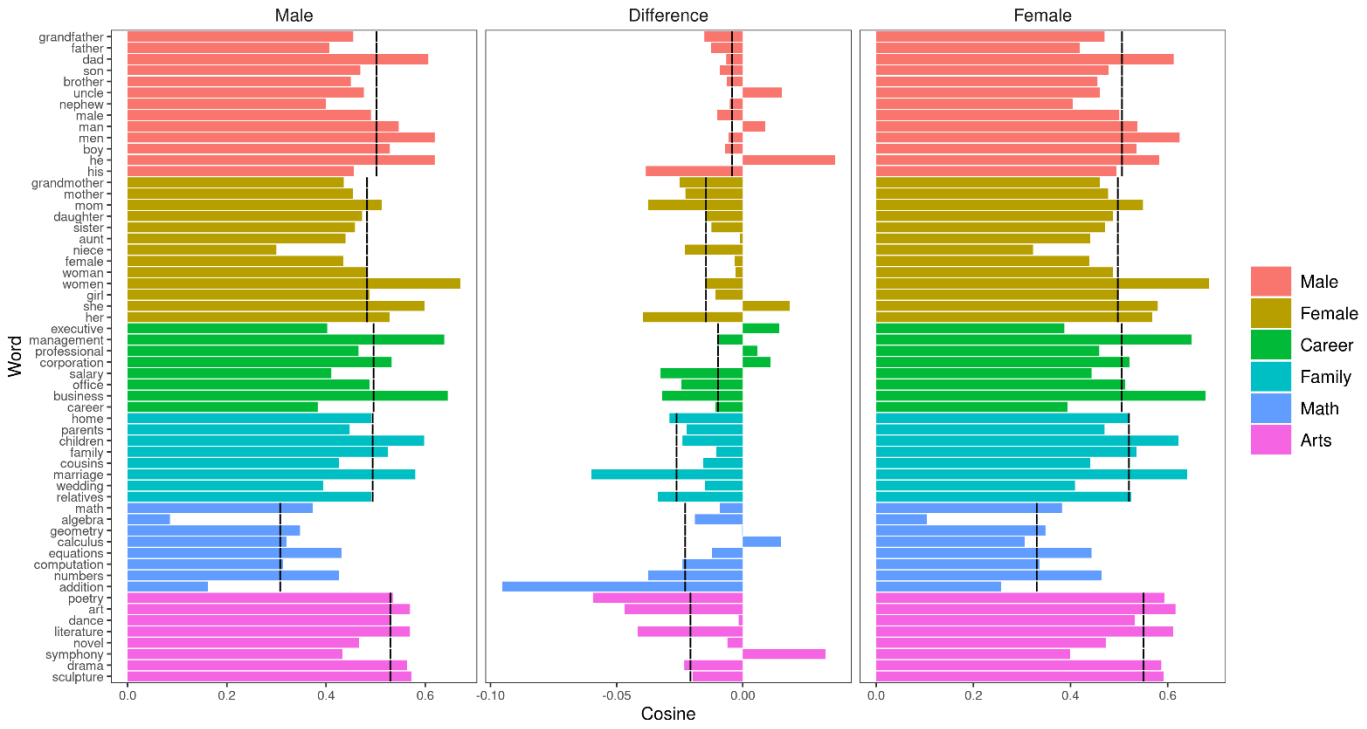
Results from experiment 6 for dominance. In the two subplots, the x axis displays the decade and the y axis displays the normalized measurement of dominance and normalized empirical measurements (war casualties, spending, and personnel). The top subplot displays the results for the Chronicling America corpus and the bottom subplot displays the results for the COHA corpus. All variables are standardized to offer easy visual comparison (i.e., all four variables are converted to z scores). The horizontal dotted line represents the z score corresponding to the center of the dominance scale (e.g., a 5 on the 9-point Likert-type scale).



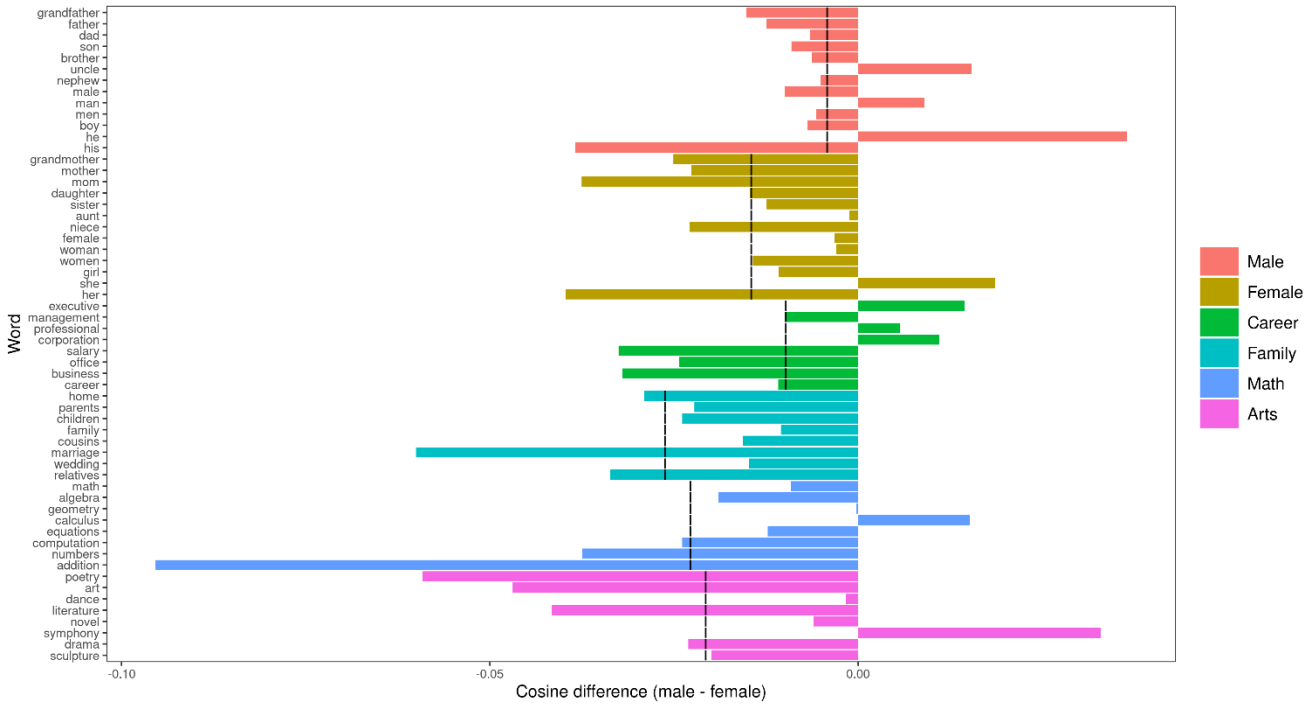
Results for experiment 7 (Chronicling America corpus only). In the left pane, the y axis shows 58 words organized according to six color coded categories (male, female, career, family, math, and arts). The x axis shows the cosine similarity between a male name concept vector and each of the 58 words. In the right pane the y axis shows the same 58 words organized according to six color coded categories (male, female, career, family, math, and arts). The x axis shows the cosine similarity between a female name concept vector and each of the 58 words. The center pane shows the same 58 words as the left and right pane but shows the difference between each pair of bars from the left and right pane (e.g., cosine(male names, grandfather) - cosine(female names, grandfather)). The vertical black lines represent the mean of each set of color bars in a given subplot. The plot displays data from the most recent data of the corpus.



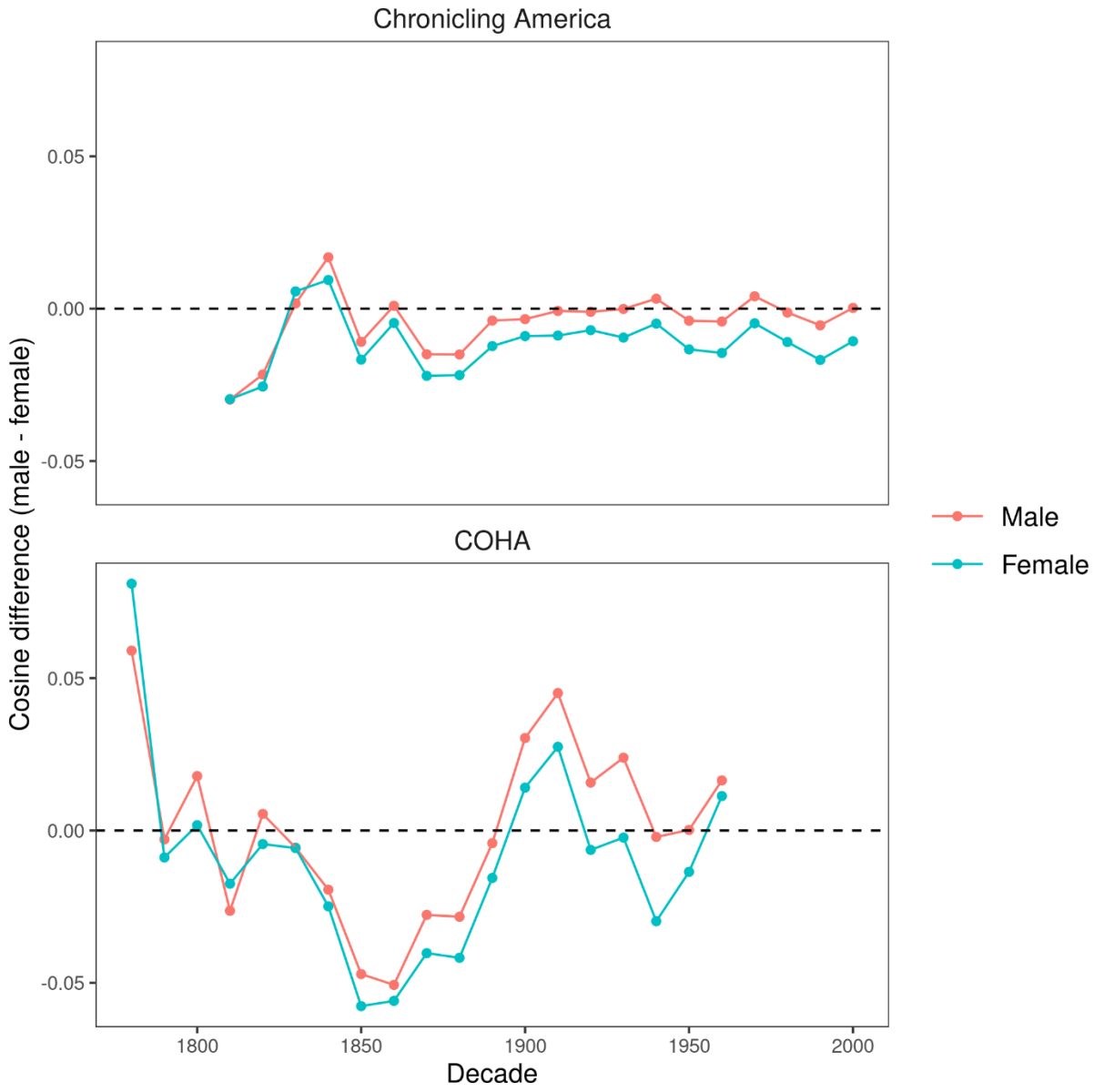
Results for experiment 7 (Chronicling America corpus only) focused on the center pane of the previous figure.



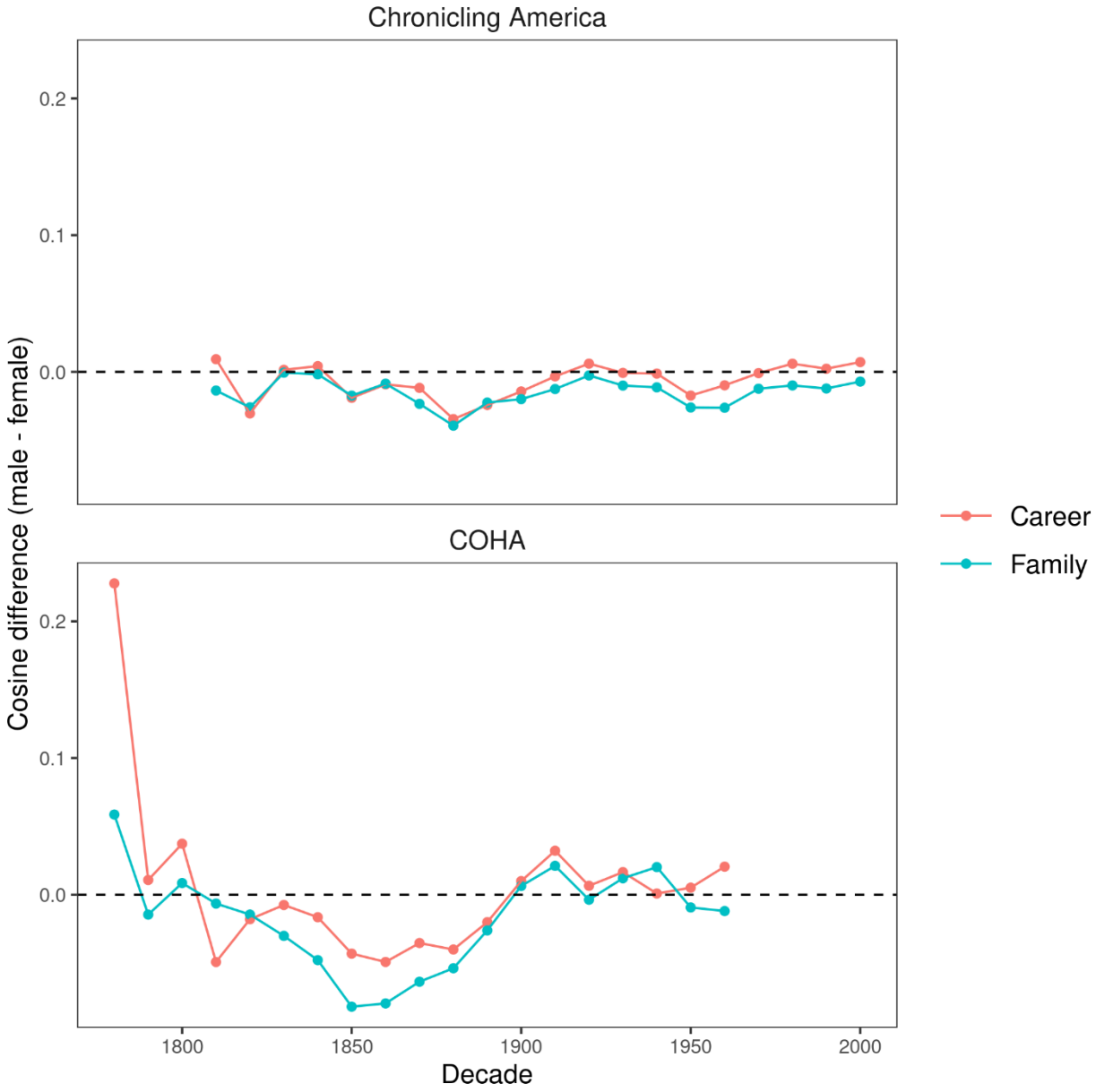
Results for experiment 7 (COHA corpus only). In the left pane, the y axis shows 58 words organized according to six color coded categories (male, female, career, family, math, and arts). The x axis shows the cosine similarity between a male name concept vector and each of the 58 words. In the right pane the y axis shows the same 58 words organized according to six color coded categories (male, female, career, family, math, and arts). The x axis shows the cosine similarity between a female name concept vector and each of the 58 words. The center pane shows the same 58 words as the left and right pane but shows the difference between each pair of bars from the left and right pane (e.g., cosine(male names, grandfather) - cosine(female names, grandfather)). The vertical black lines represent the mean of each set of color bars in a given subplot. The plot displays data from the most recent data of the corpus.



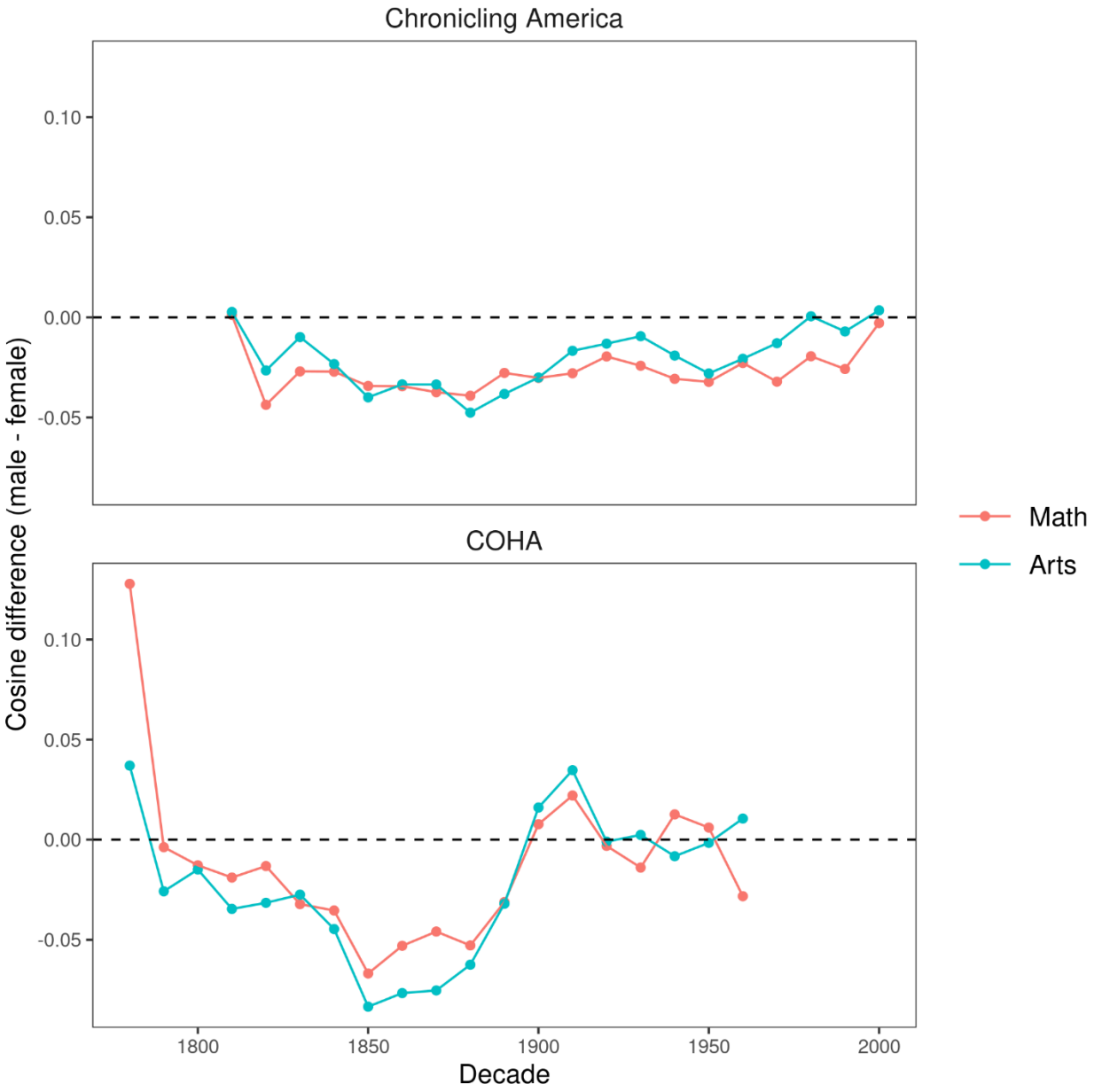
Results for experiment 7 (Chronicling America corpus only) focused on the center pane of the previous figure.



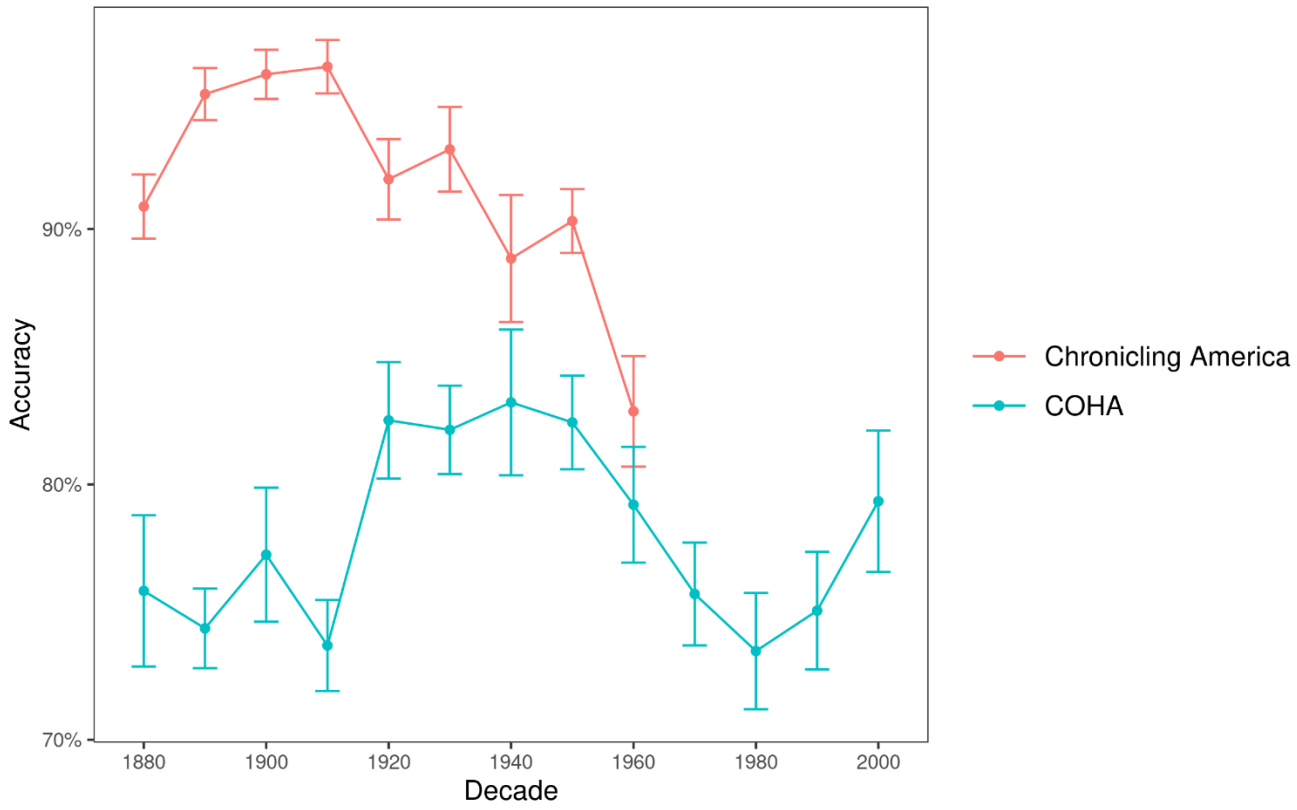
Results for experiment 7. The x axis shows decade and the y axis shows the difference in average cosine similarity between the male name concept vector and female name concept vector. The top subplot shows the results for the Chronicling America corpus and the bottom subplot shows the results for the COHA corpus. The points show the average cosine similarity difference color coded by each category of words (male, female).



Results for experiment 7. The x axis shows decade and the y axis shows the difference in average cosine similarity between the male name concept vector and female name concept vector. The top subplot shows the results for the Chronicling America corpus and the bottom subplot shows the results for the COHA corpus. The points show the average cosine similarity difference color coded by each category of words (career, family).



Results for experiment 7. The x axis shows decade and the y axis shows the difference in average cosine similarity between the male name concept vector and female name concept vector. The top subplot shows the results for the Chronicling America corpus and the bottom subplot shows the results for the COHA corpus. The points show the average cosine similarity difference color coded by each category of words (career, family).



Results of experiment 8 across all decades. The x axis shows decade and the y axis is the accuracy of the model. Error bars show standard error of the mean.