Three Families of Automated Text Analysis

Austin van Loon

Stanford University

Abstract

Since the beginning of this millennium, data in the form of human-generated text in a machine-readable format has become increasingly available to social scientists, presenting a unique window into social life. However, harnessing vast quantities of this highly unstructured data in a systematic way presents a unique combination of analytical and methodological challenges. Luckily, our understanding of how to overcome these challenges has also developed greatly over this same period. In this article, I present a novel typology of the methods social scientists have used to analyze text data at scale in the interest of testing and developing social theory. I describe three "families" of methods: analyses of (1) term frequency, (2) document structure, and (3) semantic similarity. For each family of methods, I discuss their logical and statistical foundations, analytical strengths and weaknesses, as well as prominent variants and applications.

Keywords: text as data, automated text analysis, text analysis, social science

Austin van Loon is a PhD candidate in the Stanford University Department of Sociology. His dissertation uses computational methods to examine the individual and collective dynamics of interpretation.

1 Introduction

Multiple intersecting social forces—such as increased global literacy rates (Roser and Oritz-Ospina 2016) and the advance of communication technologies (see Poe 2010)—have fundamentally altered the nature of text in society. Three such changes are worth mentioning explicitly, as they have important consequences for social science. The first is that the text-producing segment of the population has become more representative of the population as a whole. Whereas in the past only elite actors could organize their thoughts via the written word (Houston 1983), this is now a common task for most humans globally.[1] The second change is that more and more of social life happens through the medium of text. Dating happens through the exchange of messages (Rosenfeld, Thomas, and Hausen 2019), orders are given to employees through emails (Pew 2008), and politicians communicate with their constituents through posts on social media (Stier, Bleier, Lietz, and Strohmaier 2018). The final change is that naturally generated, machine-readable text is more plentiful and more easily accessible to social scientists. Books are continuously being digitized,[2] social media data can be easily accessed through application programming interfaces (APIs),[3] and even the smallest and most inexpensive of data storage devices can store hundreds of thousands of emails.[4] Given these changes, it is no surprise that social scientists have increasingly turned to the automated analysis of written language to study the goings-on of social life.

Text boasts important advantages over other forms of data used in social science. First, language is the most effective tool humans have devised for expressing ourselves, so there is little doubt that an explanation of one's attitude is more representative of their thoughts and feelings than, say, a Likert scale response (see Rocklage, Rucker, and Nordgren 2021). Second, digitally stored text exchanges are available to researchers in perfect fidelity, at least in principle. That is, if an interaction happens entirely over social media, we have access to that conversation in its entirety; there's no reduction from the phenomenon being analyzed to the data we have available. Third, because the act of text-generation is so ubiquitous, the automated analysis of large text corpora offers the possibility of studying (unobtrusively, I might add) populations that might not otherwise participate in academic research (on the challenges with securing the participation of elites in research studies, see Cousin, Khan, and Mears 2018; Mikecz 2012). Judges write decisions but are unlikely to answer a survey. CEOs write books but are unlikely to participate in experiments. Celebrities compose tweets and statuses, but they are unlikely to sit down for in-depth interviews. Harnessing this data at scale, then, presents a truly exciting frontier for developing and testing theories of the social.

However, analyzing text at scale also presents a unique combination of obstacles that, if not handled with care, could lead researchers disastrously astray. All automated text analysis relies on grossly simplifying written language, both in its representation and its interpretation. In terms of its representation, written language is an ordered sequence of incommensurate symbols (characters), meaning that even texts of relatively short length have an incredibly large number of possible

[1] For instance, The International Telecommunication Union (n.d.) estimates that 63% of the world's population used the internet in 2021, a technology that is difficult to use without a command of language

[2] See, for instance, Project Gutenberg (https://www.gutenberg.org/) and the Google Books Corpus (https://www.english-corpora.org/googlebooks/)

[3] See, for instance, the Twitter (https://developer.twitter.com/en/docs/twitter-api) and Reddit (https://www.reddit.com/dev/api/) APIs.

[4] If emails are on average 75 kilobytes, a 16-gigabyte flash drive can hold more than 200,000 emails.

realizations. For instance, a tweet of length 280 characters, of which there are 104 on the standard US keyboard, has about $6*10^{564}$ possible realizations. To represent text in complete fidelity—that is, to make no assumptions about what makes two texts similar to one another—one would need to represent each unique text as its own column in a dataset. Besides the impossibility of storing a data set with that many columns on any server (the entire internet was estimated to have about $3*10^{23}$ bits of information in 2020), such a matrix would be incredibly sparse and impractical for the purposes of statistical inference.

With respect to text's interpretation, meaning—the thing we typically want to extract from text in one form or another—is an infamously elusive construct. The meaning of various passages in The Holy Bible, which is the most read book in the world and has been widely available for at least five-hundred years (and available to some for much longer), is still hotly debated by experts who have spent much of their lives dedicated to its critical reading. The idea, then, that we could train an algorithm to perfectly extract the full meaning of a corpus of interest is doubtful on its face. The difficulty interpreting text partially arises from what might be considered structural deficiencies of human language as a medium of communication, in which words are overloaded with meaning (e.g. "tie" might refer to finishing a race at the same time, an article of clothing you wear around your neck, fastening one object to another, or, in some academic settings, a relationship between two individuals) and meanings are overloaded with reasonable linguistic representations (e.g. although the phrases "Mark Granovetter is smart" and "The author of 'The Strength of Weak Ties' possesses great intelligence" might mean the same thing for all intents and purposes, they actually have no words in common). The correspondence between language and meaning is also frustrated by complexities that are inherent to meaning—for instance that it is contingent on context, intention, and other factors typically difficult to observe in the setting of automated text analysis.

These obstacles should not dissuade social scientists from using automated text analysis to examine their important research questions. Especially in the last twenty-five years, researchers have developed incredibly powerful tools for automatically analyzing text. However, many of these approaches were developed by computer scientists and are being repurposed *ad hoc* for social science, so their links to theory building and testing are still being explored. The primary goal of this article is to distill and describe what we have learned about the links between these methods and theory building for social science. By focusing on this oft-overlooked aspect of automated text analysis methods, I hope this article has something to impart on both the uninitiated and the expert. As a secondary goal I hope to describe the intuitive foundations and popular applications of these methods.

2 A Novel Typology of Text-Analytic Methods

I divide the most common methods in automated text analysis today into three families, each of which I will review in the following sections. The first family, term frequency analysis, represents text as observations that vary in how often certain strings of characters (e.g., words) appear. The second family, document structure analysis, assumes one can extract from word co-occurrence statistics what any given document is "about" (i.e., what the appropriate keywords or themes are) and represents text as observations that vary on this feature. The third family, semantic similarity analysis, attempts to quantify the meaning of strings of characters and represents texts as collections of such meanings.

Most existing reviews of automated text analysis (e.g., Grimmer and Stewart 2013; Gentzkow et al 2019) divide methods on the basis of what role they typically serve in the research process. For instance, a method is either helpful for predicting known categories or discovering unknown categories. My typology instead divides methods based on the nature of the features they extract from text. Importantly, unlike previous typologies, one might rely on any of the families I introduce (or multiple simultaneously) to complete the same task. For instance, if one wants to take a set of labels applied to a small set of documents via hand-coding and automatically apply that same labeling scheme to a giant corpus, one might use a supervised machine learning algorithm that relies on term frequencies, document structure, and/or semantic similarities. Thus, while this typology is less practical, I believe it is more analytically insightful.

What are some of the advantages of this more analytical typology over other, more practical ones? First, thinking of methods in terms of these features helps researchers think more systematically about both operationalizing theoretical constructs in terms of linguistic features and interpreting patterns of linguistic features as theoretically meaningful. This is because these linguistic features are the actual empirical measures used in analyses—the ones that ostensibly correspond to theoretical constructs. On a closely related note, knowing that two methods are useful for uncovering unknown categories won't help the analyst make theoretical sense of discordant results from applying two methods to the same corpus—but knowing that one relies on term frequencies while the other relies on document structure might.

A second benefit of thinking about automated text analysis methods through this lens is that it will help social scientists more effectively engage with future advances in computational linguistics. Researchers in this field are constantly improving old methods and developing new ones. In fact, this is—in stark contrast to social science—the contribution of its modal paper. Thinking in terms of the typology presented here will help social scientists see the analytical commonalities between these new methods and methods with which they are already familiar. Additionally, methodological innovations vary in how important they are for how we operationalize our theories. Focusing on the analytical features of text analysis methods primes researchers to be able to distinguish between the innovations in computational linguistics that are important and relatively unimportant for social science.

I unfortunately do not have the space here to cover all the methods that belong to each family. I unfairly and somewhat capriciously decide on two sub-categories for each family, based on a combination of which are most popular in the social sciences, which effectively demonstrate the diversity and fault lines within each family, and what I am personally familiar with. In fact, the three-family typology I deploy here is not exhaustive—and as the field of automated text analysis develops, entirely new families might become dominant. Despite these caveats, I still believe that identifying such commonalities across methods at this point in the development of the field is important. I hope the way I divide methods here highlights the theoretical properties of naturally generated language and makes researchers think more deeply about what they are learning when they apply these different methods to a corpus.

3 Term Frequency Analysis

Word choice has long been central to social scientific theory. In psychology, Freud (1989 [1901]) famously argued that even unintentional word choices (a parapraxis or "Freudian slip") reveal deep aspects of the speaker's subconscious. In sociology, it has long been understood that word choice isn't only a function of the denotative meaning that is meant to be communicated but also the symbolic value of words (Bourdieu 2019 [1979]) as well as the socio-historical context in which words are uttered (Elias 1969). Such work suggests that the systematic analysis of word choice in communication can reveal important attributes of the communicator(s) as well as the context in which the communication is occurring.

Realizing this, social scientists have leveraged the increased availability of computing resources and machine-readable natural language to measure patterns in word choice—measured as how often particular terms are used—at scale. I term such methods term frequency analysis. There are broadly two categories of such methods. The first, the closed-vocabulary approach, specifies *a priori* a set of theoretical constructs to be operationalized through word choice and text metadata. The second, the open-vocabulary approach, instead inductively analyzes word choice to find patterns that explain some aspect of text metadata. I describe and provide examples of each in turn below and then compare them in the discussion section.

3.1 Closed-Vocabulary Approach

One use of term frequencies is as variables in a standard hypothesis testing framework. That is, one starts with a theory about the relationship between two or more theoretical concepts, and the (often normalized) frequency of certain terms is the chosen operationalization of one or more of those concepts. Then the frequency of those terms is used as a variable in some statistical procedure such as a regression analysis. The key assumption, often backed by some validation exercise(s), is that the prevalence of one or a set of terms meaningfully corresponds to the theorized construct.

As an example, consider the recent work of Choi et al (2022), who develop a "threat dictionary". That is, they derive a set of words (or a *lexicon*) for which they argue their presence in mass communication indicates a cultural feeling of threat. They validate this claim by showing that upticks in the prevalence of these words in US newspapers correspond to international conflicts, the spread of disease, and natural disasters in time series analyses. Having validated their measure, they test theories that relate feelings of threat to cultural tightness (also measured through language), political attitudes, and changes in macroeconomic conditions.

There are far too many such dictionaries to fully enumerate here, but a few are worth mentioning briefly. Gaucher, Friesen, and Kay (2011) develop a dictionary of "masculine" and "feminine" words to track gendered language in job advertisements. The moral foundations dictionary—originally developed by Graham, Haidt, and Nosek (2009; but see moralfoundations.org for updated material)—tracks the kinds of moral intuitions authors deploy. Nicolas, Bai, and Fiske (2020) develop a set of lexica to track the use of stereotypes along with a general algorithmic approach for building dictionaries to tap theoretical constructs of interest.

There is no single, agreed-upon way to arrive at a set of words whose use corresponds to a theoretical construct, though a roughly common framework can be observed from the work reviewed here. First, one develops a set of "seed words", or keywords that tightly correspond to the concept of interest. Sometimes this is just the word for to concept(s) of interest (e.g., "threat" in Choi et al 2022 or the names of the moral foundations in Graham, Haidt, and Nosek 2009), and sometimes it involves drawing on previous literature (as in Gaucher, Friesen, and Kay 2011; Nicolas, Bai, and Fiske 2020).

Then, these seed words might be expanded upon, either by human judgement (Graham, Haidt, and Nosek 2009), word embeddings (Choi et al 2022; see the "semantic similarity analysis" section for more details on word embeddings), or other tools such as WordNet (Nicolas, Bai, and Fiske 2020)—a database of English concepts and conceptual relationships curated by experts (Miller 1995). There may also be a pruning phase where words that are deemed too distantly related to the concept(s) of interest are removed (Graham, Haidt, and Nosek 2009; Choi et al 2022). These steps may be repeated multiple times until a satisfactory lexicon is achieved (Nicolas, Bai, and Fiske 2020).

Similarly, there is no single way to validate one's lexicon after it has been compiled, but the three primary ways this is done to date are represented by these same articles. One might, as Choi et al (2022) did, demonstrate convergent validity by showing that variation in their lexical measure correlates with events that are plausibly associated with their theoretical construct. One might also, as Nicolas, Bai, and Fiske (2020) did, compare their lexicon to words elicited directly from respondents in a survey. Finally, one should always, as do Graham, Haidt, and Nosek (2009), go back to specific examples words in their lexicon being used in their corpus to confirm they are being used as expected. If feasible, one should do each of these steps or some variation of them. Additionally, however, it would behoove researchers to seek to demonstrate divergent validity—that their lexical measure is largely uncorrelated with related but distinct theoretical constructs—along with convergent validity.

Linguistic Inquiry and Word Count or LIWC (Pennebaker, Booth, and Francis 2007; Boyd et al. 2022)[5] is what one might call a "general purpose" dictionary[6] as well as proprietary software to analyze text files using it—we'll focus on the former. The dictionary is made up of sets of words and word stems (e.g., "hungr*", which corresponds to "hungry", "hungriest", etc.) that purportedly tap into distinct, hierarchically organized constructs (e.g., 1st person singular pronouns are a subset of personal pronouns which are themselves a subset of function words). The process by which words are included in these sets has changed over time along with new iterations of LIWC, but it is generally an iterative process leveraging dictionaries and thesauruses, natural language processing tools, human judgements, and psychometric evaluation (Pennebaker, Boyd, Jordan, and Blackburn 2015; Tausczik and Pennebaker 2010).

First used to study the relationship between descriptions of traumatic events and health (Pennebaker 1993; Pennebaker and Francis 1996; Pennebaker, Mayne, and Francis 1997), LIWC has been used for over a quarter of a century by social scientists to study a wide range of phenomena. Stirman and Pennebaker (2001) show that poets who eventually commit suicide use more individual-centric language in their poetry. Ashokkumar and Pennebaker (2021) analyze how the frequency of terms in many LIWC categories change over the course of the COVID-19 pandemic, inferring changes in macro-psychological conditions. Bail, Brown, and Mann (2017) use the prevalence of "emotion" words and "cognition" words to measure discursive currents in the field of advocacy organizations.

Researchers have also thought of creative ways to use LIWC to measure more nuanced theoretical constructs. Ashokkumar and Pennebaker (2022), for instance, validate a measure of group identity strength that is equal to the ratio of the frequencies of "we" words to "cognition" words occurring in text. Brady et al (2017) use the intersection of the moral foundations dictionary and "emotion" words

[5] See see https://www.liwc.app/ for more information.
[6] DICTION (Hart 1984) is another example. See https://dictionsoftware.com/diction-overview/ for more information.

from LIWC to measure the moral-emotional strength of social media posts, showing that this is correlated with how many times they are shared (but see Burton et al 2021). Goldberg et al. (2016) use divergence in the probability distributions over many LIWC categories between an individual and their interaction partners in work emails as a measure of situated cultural fit.

3.2 Open-Vocabulary Approach

Another approach to the analysis of term frequencies is entirely inductive, allowing for interesting relationships between term frequency and metadata to emerge from the corpus. This approach can be seen as analogous to gene-wide associations studies (GWAS; see Tam et al 2019), testing the association between many term frequencies and a pre-specified variable of interest. As with GWAS, interpreting results of an open-vocabulary analysis should be done with caution.

In differential language analysis (DLA), the analyst selects a dependent variable—often document-level metadata—that is of interest. They then identify a vocabulary that well-represents the corpus being analyzed (often the *K* most common terms in the corpus excluding stop words).[7] Then, for each term in the vocabulary, the association (e.g., the Pearson correlation) between its prevalence and the dependent variable is calculated. The *P*-values from each test is then corrected for multiple comparisons, reducing the risk of false discovery. Finally, the results are summarized (for instance graphically in word clouds) and presented to the researcher, who then derives insights about the empirical phenomenon of interest. These insights rely on interpreting *a posteriori* what theoretically interesting concepts are represented by the term frequencies, which might require additional analyses to justify and defend.

While several prior studies had already used open-vocabulary analysis (e.g., Monroe et al 2009; Yarkoni 2010; Einstein et al 2011), Schwartz et al (2013) introduced the methodology of DLA to examine the linguistic features most correlated with individuals' gender, age, and personality traits. With respect to personality, some of their findings accorded with existing theory while other results suggested entirely novel theoretical directions. Researchers have also shown that these methods can be harnessed to make socially important predictions. Eichstaedt et al (2015) show that open-vocabulary analysis of the social media discourse in communities across the U.S. can be used to predict the prevalence of heart disease therein. Eichstaedt et al (2018) find that DLA can be used to predict whether someone has depression with the same accuracy as clinical evaluations, and provide novel insights into the emotional, interpersonal, and cognitive antecedents of depression.

3.3 Discussion

The open-vocabulary and closed-vocabulary approaches to term frequency analysis have unique strengths and weaknesses. Jaidka et al (2020) show how dictionaries like LIWC, because they are not data-driven in the same way as DLA, can suffer more from issues associated with term ambiguity. For instance, they show that many words contained in the "positive emotions" dictionary of LIWC (e.g. "love", "lmfao", "respect") are negatively associated with community-level happiness and many words in the "negative emotions" dictionary (e.g. "dominate", "critical", "hungover") positively correlate

[7] The estimated prevalence of various topics (see the document structure analysis section) in each document are sometimes used instead of term frequencies.

with community-level happiness. This could lead researchers astray if they dogmatically assume that a LIWC dictionary directly corresponds to their theoretical construct of interest.

Eichstaedt et al (2021) show that the open-vocabulary approach generally outperforms the closed-vocabulary approach when predicting psychologically important variables of texts' authors. However, social scientists are often not interested in prediction but in testing existing social theory, which DLA is unambiguously inefficient in doing. For instance, while the use of moral-emotional language might not explain a high proportion of the variance in how often a tweet is retweeted (and therefore might not surface in a DLA analysis), we might have theoretical reasons be interested in whether they are reliably related. Constraining the set of analyzed linguistic features to those that plausibly correspond to the construct(s) of interest *a priori* and allowing them to load onto a single meaningful factor allows for a more efficient test of a theory. However, the rigorous and insightful comparisons of Jaidka et al (2020) do highlight the importance of validating one's operationalization of theoretical constructs through term frequencies.

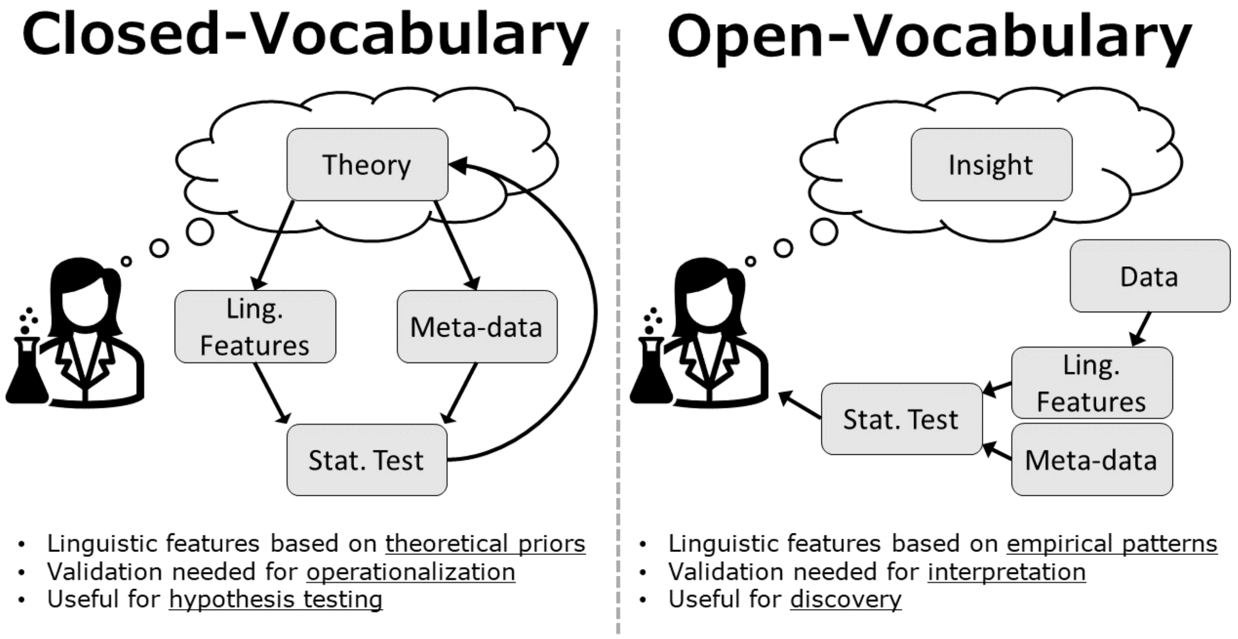
Closed-Vocabulary
Open-Vocabulary
Theory
Ling. Features
Meta-data
Stat. Test
Insight
Data
Ling. Features
Stat. Test
Meta-data
Linguistic features based on theoretical priors
Validation needed for operationalization
Useful for hypothesis testing
Linguistic features based on empirical patterns
Validation needed for interpretation
Useful for discovery

Figure 2. A conceptual figure depicting prototypical closed- (left) and open- (right) vocabulary term frequency analysis. In closed-vocabulary analysis, the researcher derives from theory a set of linguistic features and metadata that operationalize the relevant constructs and tests their relationship via an appropriate statistical test. The results of that test speak to the soundness of the theory. In open-vocabulary analysis, linguistic features are automatically derived from the data, and each is tested for association with metadata via a statistical test. The results of these statistical tests are summarized for the researcher, from which she derives insights about the data.

The core assumption of term frequency analyses is that a term's prevalence consistently reflects something meaningful about the document, its author, or the context in which the document was produced. As discussed in the introduction of this article, however, we know that words have multiple meanings, and this can lead researchers astray. This is especially problematic for general-purpose dictionaries like LIWC, which expect words to mean the same thing not only across usages within a corpus, but also across all corpora to which it might be applied. All term frequency analyses, whether they are open- or closed-vocabulary, ultimately rely on researcher-determined validations to justify the correspondence between term frequency and theoretical constructs. Stronger norms around how to determine this correspondence would greatly help the field of automated text analysis advance and produce more reliable scientific insights.

4 Document Structure Analysis

While term frequency analyses generally treat terms as atomistic, documents are typically the unit at which text is meaningfully engaged with in everyday life. We talk more often about specific tweets, policy platforms, and text messages than how often an individual uses a certain word or class of words. The second family of methods I discuss analyzes patterns at the level of the document, seeking to estimate hidden patterns in the way words are distributed amongst them. I call this family of methods document structure analysis. Arguably the first method for automated document structure analysis was latent semantic indexing/analysis, which was introduced by Deerwester et al (1990) and quickly found its way into Psychology (Landauer and Dumais 1997).

One way to think about this family of methods is that they assume there are a set of latent variables, often called "topics", for which their values determine the content of a particular document. The values of these variables correspond to what each document "is about". For instance, a document might be highly "political", and will therefore contain a high proportion of words related to politics. That same document might not be "pop-cultural" and will therefore not contain many words related to pop culture. These methods inductively arrive at the set of "things" documents in a given corpus are "about," and then determine what each document therein is "about". By estimating these latent variables, these methods (either explicitly or implicitly) posit an analyzable structure to the complex process by which documents are generated.

An alternative way to think about these methods is that they inductively reduce the dimensionality of the language space in which documents are represented (like PCA for structured data) and can be motivated by recognizing that most socio-linguistically interesting features of text emerge not from the use of single words but from the co-presence of multiple words. For instance, the presence of the word "March" in a document tells you one thing when paired with "Soldier", "Formation", and "Uniform", but tells you something entirely different when paired with words like "January", "February", and "April". In this case, "topics" can instead be thought of as the clusters of words whose combined presence meaningfully divide the documents.

The role of document structure analysis in theory building and testing, then, can be one of at least two things. First, it can be used to aid grounded theory (Glaser and Strauss 1967) or other inductive approaches by highlighting important themes in a corpus (see Nelson 2020). For instance, if one is interested in the sociology of knowledge, then finding the core themes in a corpus of academic writings during the scientific revolution (or how they change over time) might provide novel theoretical insights that could be built into a theory. One might even take an abductive approach (Tavory and Timmermans 2014), testing emerging theories in the same data (ideally in a held-out partition; see Egami et al 2018). Alternatively, the topics present in documents can be used as measures that summarize the content of a corpus for deductive tests of pre-specified theories. For instance, a theory that implies climate change discourse drastically changed around a certain event might be tested by studying the changing prevalences or contents of topics around that event.

Though other approaches exist—such as non-negative matrix factorization (Lee and Seung 1999; Pauca et al 2004)[8], probabilistic latent semantic analysis (Hoffman 2013)[9], and Doc2Vec (Le and Mikolov 2014)—there are two dominant approaches to document structure analysis. The first is a set of methods which infer topics through Bayesian inference and are what are widely referred to as "topic models". The second set of approaches treat the document-term matrix (or a transformation of it) as an adjacency matrix, which is then modeled as a network. Community detection algorithms are then deployed to identify topics. I will now discuss each approach in turn.

4.1 Bayesian Approaches

Outside of sociology, Bayesian inference is by far the most popular way to estimate topics in a corpus. The foundational model is latent Dirichlet allocation (LDA; Blei et al 2003), named after the probability distribution used as a prior both for the distribution of topics in documents and the distribution of words' prevalences within topics.[10] In this approach, a "topic" is modeled as a multinomial distribution over all terms in the vocabulary. For instance, in one topic words such as "election", "president", and "government" might be assigned high probabilities relative to other words like "baseball", "retweet", and "gourmet". This might be called a "political" topic.

The model is generative, meaning it assumes a stochastic process by which corpora are generated and then—when applied to a corpus—fine-tunes this process such that it maximizes the likelihood of that corpus resulting from it. Although it is highly artificial, it is helpful to first approach this model by walking through this assumed data-generating process conceptually. First, an author selects the length of their document (N) as well as the proportion of the document that is expected to be made up of each of K (a pre-specified number) of topics (θ). Then, for each word apportioned to a topic, a word is chosen from the vocabulary on the basis of how likely words are to occur in that topic (φ). This results in a "bag-of-words", which can be represented in a document-term matrix just like any real document. See Figure 3 for a representation of this process.

The model is fit to empirical data via Bayesian inference. Specifically, each step of the data generation process is modeled as a draw from a distribution. The researcher sets an expectation (or prior) for what these distributions might look like, which is then updated based on the observed data.[11] N (a count) is modeled as a draw from a Poisson distribution (ξ), while θ and φ (both multinomial distributions) are modeled as draws from two separate Dirichlet distributions (α and β, respectively). The priors on α and β are usually set such that values of θ and φ with concentrated probability masses are more likely than ones with evenly distributed probability masses. Importantly, this means that

[8] See de Paulo Faleiros and de Andrade Lopes (2016) for more on the relationship between LDA and non-negative matrix factorization.

[9] See Girolami and Kabán (2003) for more on the relationship between LDA and probabilistic latent semantic analysis.

[10] The specifics are beyond the scope of this article, but in brief the Dirichlet distribution—described by K (a pre-specified number of) parameters—can be thought of as describing the probability of all possible multinomial distributions of K random variables (like how a normal distribution describes the probability of all possible real numbers). Importantly, different values of the distribution's parameters can express how "concentrated" we expect a given multinomial distribution to be. When all parameters of a Dirichlet distribution are equal, setting their values to be greater than one makes multinomial distributions where probability is spread over all variables more likely while setting their values to be less than one favors distributions where any one of the K variables is more probable than the others.

[11] This is typically done either via variational inference (Blei et al 2016) or Gibbs Sampling (Griffiths and Steyvers 2004; Gelfand 2000).

LDA roughly seeks to estimate the prevalence of topics over documents and words within topics such that (a) documents are mostly made up of a small number of topics and (b) topics are mostly made up of only a small number of words.

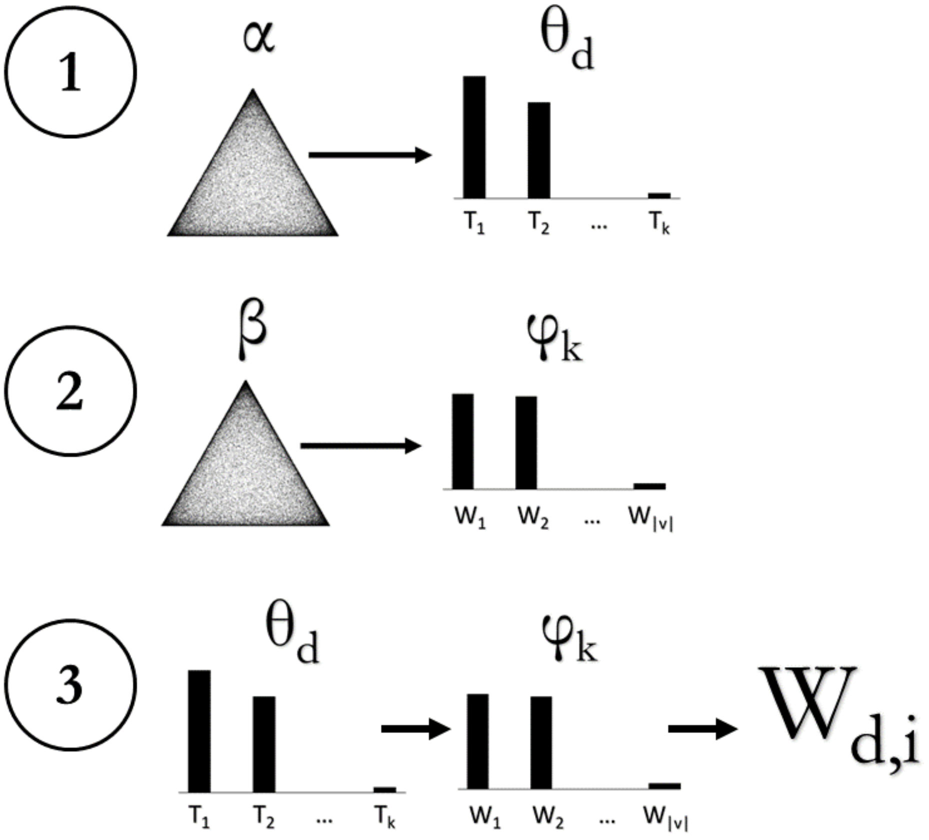

Figure 3. A depiction of the data generation process assumed by LDA. (1) for each document, a multinomial distribution θ_d (representing the proportion of words in the document expected to be from each topic) is drawn from a Dirichlet distribution α. (2) for each of the *K* topics, a multinomial distribution (representing the probability of a word from a topic being equal to any given word in the vocabulary) φ_k is drawn from a Dirichlet distribution β. (3) for each word in each document $W_{d,i}$, a topic *k* from θ_d is drawn, and then a word from φ_k is drawn.

DiMaggio et al (2013)[12] apply LDA to nearly eight-thousand newspaper articles to abductively study the politicization of the arts. They explore the topics that result from their LDA model to characterize the important themes in their discourse, and then iterate between quantitative analysis and deep reading to examine how the prevalence and content of these topics vary over outlet and time. Such an abductive approach is the overwhelmingly common way to use topic models in the social sciences. Hackett et al (2021), as an example of using topic modeling for deductive analysis, use LDA to reduce

[12] It is worth mentioning that this article is one in a special issue of *Poetics* dedicated entirely to topic models (Mohr and Bogdanov 2013), which is a great place to find examples of LDA being used for sociological analysis.

the dimensionality of the language space in the service of testing hypotheses related to the origins of scientific synthesis. For a particular scientific article, they quantify its "topical diversity" by examining the share of different topics estimated to be in its title and abstract. They then calculate this for each article in their corpus and test whether it covaries with metadata of interest.

A popular variant of LDA is the correlated topic model (CTM; Blei and Lafferty 2006). The key innovation of the CTM is to explicitly model the fact that topic prevalence's might be correlated across documents. This is achieved by using a logistic normal distribution for α (the prior for the distribution of topics over documents) instead of the Dirichlet.[13] Insofar as the prevalence of topics are "actually" correlated with one another, a CTM should improve the quality and efficiency of topic estimation over LDA. Alvero et al (2021) use a CTM to show how college applicants from different socio-economic backgrounds write systematically different college admissions essays. They further show that essay content predicts socio-economic background better than standardized test scores. Note that in this application, the theory—students from high socio-economic backgrounds will write about different topics in college admissions than students from low socio-economic backgrounds—posits that there will very much be correlations among topics, so the use of the CTM is highly appropriate.

Another popular variant of LDA that has been more recently developed is the structural topic model (STM; Roberts et al 2013; Roberts et al 2014). A generalization of other variants such as dynamic topic models (Wei et al 2012) and author topic models (Rosen-Zvi et al 2012), an STM is a CTM that incorporates the general intuition that document metadata (e.g., the year or country in which it was written, the political party of the author, the medium through which it was originally distributed) might influence the document generation process. Specifically, documents with different levels of categorical metadata can be allowed to vary either in θ (the prevalence of topics within a document) or φ (the prevalence of words within a topic, or the "content" of a topic). This is argued to improve topic estimation (Roberts et al 2016) and allows analysts to examine differences qualitatively and quantitatively. This topic modeling variant has achieved great popularity in Sociology (see, e.g., Bail et al 2017; Nelson 2021; Heiberger et al 2021) and Political Science (see, e.g., Kim 2017; Kim 2018).

4.2 Relational (Network) Approaches

Perhaps the strongest critics of these Bayesian approaches are a group of (largely) sociologists who analyze language as a fundamentally relational phenomenon, and therefore apply the toolkit of network science to the analysis of document structure. There are of course different variants of such an analysis, but the one most relevant to this article constructs a network in which nodes are words and connections between them represent how often these words occur in the same documents. Well-understood community detection algorithms can then be applied to this network to formally derive "topics", or communities of words that tend to co-occur. Finally, documents can be re-interpreted as collections of different themes in varying quantities. This approach not only offers an aesthetically pleasing and transparent way to present the content of topics (the constructed network), but also allows for intuitive analysis.

[13] Like a multivariate normal distribution, the parameters that characterize a multivariate logistic normal distribution include a covariance matrix. Although the full details are beyond the scope of this article, the logistic normal distribution is not a conjugate prior of the multinomial distribution (i.e., α is no longer "of the same form as" θ), so inference is more complex.

Rule et al (2015) create such a network from state of the union speeches delivered by American presidents. Specifically, they connect words if they occurred in the same paragraph of the same address. They then identify clusters of words, identify what each cluster substantively represents, and trace how these clusters are used over time, arguing that the modern American political consciousness arose sharply after World War I. Hoffman et al (2018), instead, map the connections between words in the Bible, which they pair with data on which preachers used which verses in sermons in the 17th and 18th centuries to study cultural contention. Hoffman (2019), instead of making a network of words, maps a network of documents (books checked out of the New York Society Library in the late 18th and early 19th century) based on how similar the words they use are. He pairs this with data about who checked out which books to study changing political alignments and material culture in early American history.

In each of these examples, the researchers began with a bimodal network (documents were connected to words) and then projected this into a one-mode network, for instance by measuring the cosine similarity between rows or columns of the adjacency matrix. Melamed (2014) warns that such an approach can be inaccurate when different numbers of communities exist among modes, proposing instead an iterative, dual-projection approach. Gerlach et al (2018) develop a promising alternative that applies hierarchical stochastic block modeling (a generative model of community detection) directly to the bimodal network represented by the document-term matrix. A major benefit of this approach is that it analytically determines an appropriate number of clusters.

It is worth mentioning that the application of network analysis to text does not necessarily imply an analysis of document structure. Indeed, the earliest instances of automated semantic network analysis attempted to get at relationships between words that were more complex than co-occurrence in documents. Danowski (1993), for instance, connects words in a network if they co-occur within a sliding window. Map analysis (e.g., Carley 1994) connects concepts with heterogeneous ties if they appear in a shared statement (e.g., "Mark Granovetter has no critics" produces a negative tie between "Mark Granovetter" and "critics"). Some contemporary work in cognitive science (see Lynn and Bassett 2020) also produces and analyzes networks from language, but on the basis of transitions between words (e.g. "text analysis is fun" would encode the following directed ties: "text" → "analysis", "analysis" → "is", and "is" → "fun"). Clearly these approaches also produce insights important for social science—but they are only connected to relational approaches to document structure analysis superficially via their use of the toolkit of networks to study natural language.

4.3 Discussion

Document structure analysis relies on the assumption that co-occurrence statistics, and therefore the boundaries of documents, are meaningful. This is fundamentally different than term frequency analysis, where terms could be arbitrarily shuffled between documents with identical metadata. The ostensible benefit of this additional assumption is that it reduces concerns related to term ambiguity by relying on term co-occurrences to detect themes within the corpus. Importantly, the reliance on co-occurrence means that decisions about the boundaries of documents are extremely important. Is the correct way to model the contents of a corpus of books to make each book a document? Each chapter? Paragraph? Sentence? The answer depends on the research question as well as particulars of the corpus. However, one guiding insight is that most document structure analysis methods assume a sparse relationship between topics and documents—that is, that documents only cover one or a few topics. This lead Hoffman et al (2018), for instance, to consider a sliding window of five verses in the Bible to be the appropriate unit of text to be considered a document (p. 95).

One of the primary weaknesses of existing approaches to document structure analysis is the reliance on difficult-to-justify researcher decisions. For instance, when performing an LDA the researcher must set priors on the distribution of topics over documents and the distribution of words within topics and deterministically set the number of topics in the corpus. Although guidelines and empirical approaches exist for making such decisions (Taddy 2012; Bischof and Airoldi 2012; Minmo et al 2011; Wallach et al 2009; Chang et al 2009), none are standard and in practice the decision is ultimately left up to what produces topics that are theoretically useful (Grimmer and Stewart 2013). Relational approaches can be better in this respect, though the decisions of what clustering algorithm to perform to identify topics, how to define the boundaries of documents, and how many words should be included in the final vocabulary still haunt this approach.

While the inductive derivation of topics is considered an analytical strength (see DiMaggio et al 2013), it can also be a practical weakness. If researchers come into the research setting with *a priori* expectations about what the core themes are in a corpus, there is no guarantee those themes will show up in most document structure analyses. Fligstein et al (2017) use LDA to analyze transcripts of meetings of the Federal Reserve, arguing that the frames they used to think about and discuss the economy stopped them from foreseeing the 2008 financial crisis. Their deductive analytical approach relied on an LDA model to capture frames related to macroeconomics and the "concepts and tools of finance and banking". They had strong theoretical priors that these were important aspects of how the Fed discussed the economy, but this alone doesn't suggest there will be distinct topics representing these constructs. In fact, the ubiquity of a frame is a reason one might *not* expect a corresponding topic to be produced from LDA, since the algorithm searches for linguistic features that *differentiate* documents. For instance, in a corpus of biology article abstracts, one should not expect a "biology" topic to emerge from an LDA analysis. If researchers vary their pre-processing and other decisions (e.g., the number of topics in LDA) until they find the themes they expected *a priori*, then the estimated topics might no longer faithfully reflect the data. Generally, it would be advisable to be open to what the important themes are of a text when using document structure analysis. If there are strong priors about the themes that are important to a corpus, one might look for special varieties that can reflect such priors, for instance the supervised topic model (Blei and McAuliffe 2007) or the seeded topic model (Lu, Ott, and Tsou 2011).

5 Semantic Similarity Analysis

Social scientists have long sought a quantify the meaning cultures associate with concepts. Charles E. Osgood, for instance, implemented surveys in 25 countries and argued that affective meaning in humans could be distilled into three basic factors: evaluation, potency, and activity (Osgood 1971). Social scientists have continued in this vein, developing new typologies and methods for understanding how individuals and groups interpret concepts that drive the goings-on of social life (see Aceves and Evans 2022). In the final family of automated text analysis I introduce, semantic similarity analysis, concepts or words used in a corpus are given a quantitative representation corresponding to their respective meaning, which are then compared.

In principle, this could be done with a variety of tools. For instance, correspondence analysis can be applied to a document-term matrix to place words in a multi-dimensional space, in which their position indicates their social meaning (see, e.g., Lebart, Salem, and Berry 1997). Language models commonly used for document structure analysis—such as topic models or semantic networks—could also produce term-specific measures (e.g., a vector of a term's probability of appearing in each topic of a corpus) that could operationalize each term's meaning. However, the method that is overwhelmingly used in the social sciences to quantitatively represent terms' meanings is so-called "word embeddings". Due to the tool's overwhelming prominence and its relative technical complexity, I'll focus on explaining the intuition behind these methods and how they can be used for semantic similarity analysis.

The distributional hypothesis (Harris 1954) states that "you shall know a word by the company it keeps" (Firth 1957). It summarizes the argument of distributional semantics, perhaps most famously championed by Wittgenstein (2010 [1953]): "the meaning of a word is its use in the language" (p. 20). The general argument (or at least the parts relevant to semantic similarity analysis) can be condensed to be that two words' "meanings" are more similar when the contexts in which they are used are more similar. Said differently, two terms are completely synonymous if they are completely interchangeable—the less interchangeable they are, the less synonymous they are.

Consider three words: "doctor", "dentist", and "orthodontist". To operationalize how interchangeable they are, we might measure how often each word is used across a set of possible contexts. Imagine, for the sake of simplicity, that there are only three linguistic contexts (ignoring other contextual factors) in which these words are used: (1) "Is there a ___ on the plane?"; (2) "The ___ will see you now"; and (3) "A ___ can fix your teeth". In the first and second contexts, "doctor" is much more likely to be used than either "dentist" or "orthodontist". In the third, "dentist" and "orthodontist" are much more likely to appear than "doctor" and are similarly likely to appear themselves. This information is plotted two different ways in Figure 4. In this hypothetical example, "dentist" and "orthodontist" being used in more similar contexts to each other than to "doctor" suggests that "dentist" and "orthodontist" are more synonymous with each other than they are with "doctor".

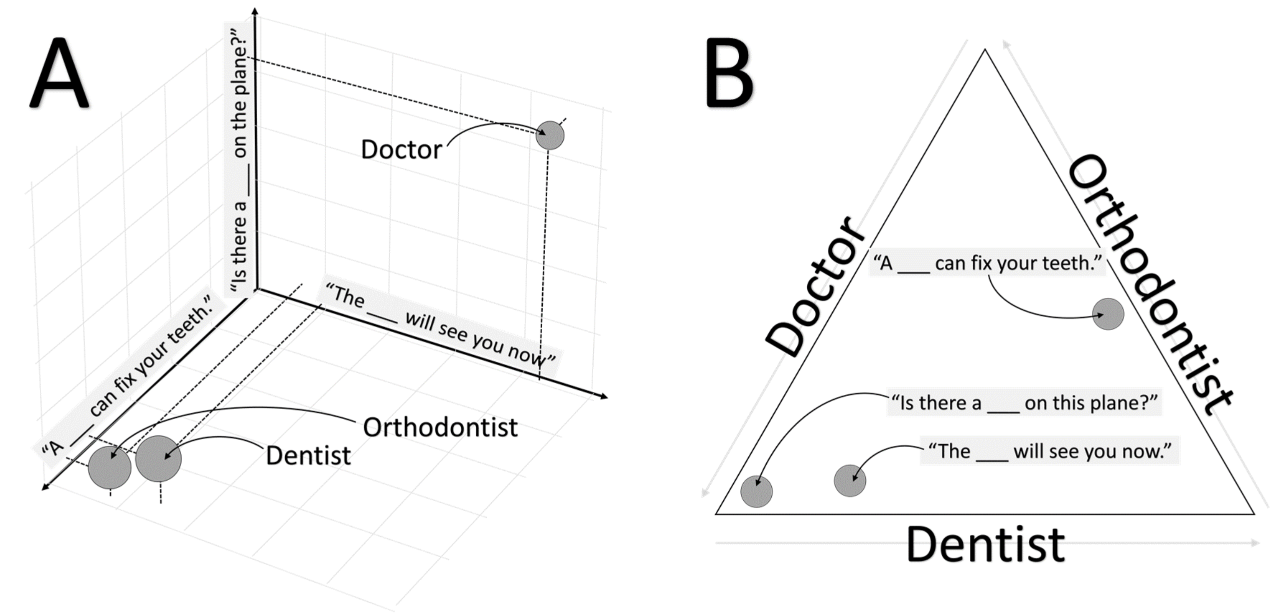
A
"Is there a ___ on the plane?"
Doctor
"The ___ will see you now"
"A ___ can fix your teeth."
Orthodontist
Dentist
B
Doctor
Orthodontist
"A ___ can fix your teeth."
"Is there a ___ on this plane?"
"The ___ will see you now."
Dentist

Figure 4. Conceptual diagram illustrating the core idea of distributional semantics. (A) three words plotted in a three-dimensional space corresponding to how often they might appear in three contexts. (B) Three contexts plotted in a simplex based on what proportion of instances might contain one of three words. If all contexts and words are represented simultaneously, the two representations are equivalent. Words are more similar in meaning in (A) the closer they are to each other and in (B) the more correlated the positions of points are on their respective axes.

Word embedding algorithms attempt to quantitatively represent this information for all words across all immediate linguistic contexts in a corpus. Word2Vec (Mikolov et al 2013) is a popular word embedding algorithm that accomplishes this with deep learning.[14] A simplified conceptual diagram of how Word2Vec (or at least one variant of it) learns this information is presented in Figure 5. An artificial neural network is constructed with an input layer (encoding the independent variables) and output layer (encoding the dependent variable) of size *V*, with each neuron in both layers corresponding to a word in the vocabulary. Then, for each sequence of words in the corpus (e.g., "The doctor will see you now" → "doctor see now"), a training example for the network is constructed where each word should be predicted from the words present in its context (e.g., "doctor" should be predicted from the presence of "see" and "now, "see" should be predicted from "doctor" and "now", etc.). For each example, the network learns to weight the connections between neurons so that it can correctly predict words from contexts. Then, an internal representation of the context resulting in the network most confidently predicting a word is taken to be a vector representation of the word itself (also known as its "embedding" or "distributed representation"). One can then think of the vectors of each word representing its position within a "semantic space": the words the neural network would predict as appearing in similar contexts, then, will have more similar representations, and therefore be closer in this space.[15]

[14] For an excellent introduction to deep learning that doesn't assume any mathematical or programming background, I highly recommend the "Deep Learning" series on YouTube from 3Blue1Brown.

[15] GloVe (Pennington et al 2014) is another popular (and potentially more fruitful for social science) variant of static word embedding algorithms, but for brevity I don't review it here as Word2Vec is more helpful for understanding these methods conceptually.

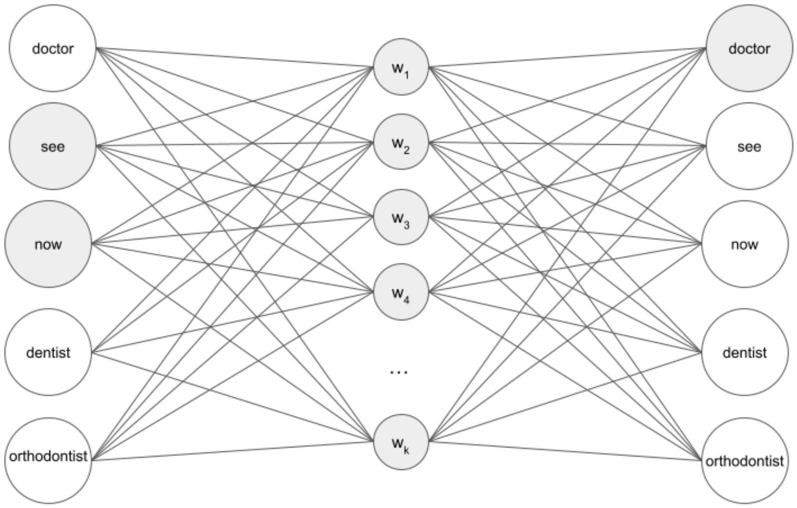

Figure 5. A conceptual diagram of how Word2Vec learns distributed representations of words' meanings from the example "The doctor will see you now".

Contextual embeddings are a recent advancement in word embeddings that rely on more complex deep learning algorithms to incorporate a word's immediate linguistic context into its vector representation. For instance, the word "ties" in "the strength of weak *ties*" and "Mark Granovetter doesn't wear *ties*" are allowed to have distinct representations. The most popular version is BERT (bidirectional encoder representation from transformers; Devlin et al 2018). Many state-of-the-art algorithms for solving natural language processing tasks (e.g., machine translation, coreference resolution, word sense disambiguation) rely on vector representations acquired via BERT or a close variant. Kjell et al (2022) even find that BERT-based representations of open-ended survey responses measure underlying psychological constructs at least as reliably as traditional, closed-form survey measures.

The comparison of term-level information (for instance their embeddings) to learn about how these symbols are used within a corpus is what makes semantic similarity analysis unique. While the use of semantic similarity analysis within the social sciences is still evolving, applications can be differentiated by whether the quantities of interest are the proximities of words within a single semantic space or differences in the same distance across semantic spaces estimated from multiple corpora. I call the former the "within-corpus approach" and the latter the "between-corpus approach". I describe each in turn below and then turn to some issues with word embeddings specifically in the discussion.

5.1 Within-Corpus Approach

Standard word embeddings, like topic models, are unsupervised—any patterns they learn are derived from the data and not expectations the analyst might have. Researchers have reasoned, then, that if psychological and cultural factors shape the process by which words come to have stable meanings in a population, then word embeddings might allow us to reverse-engineer and study those social factors via their residue in natural language. For instance, if a culture heavily stigmatizes one identity and valorizes another, this might influence the way linguistic markers of those identities are used in everyday language within that cultural context. In a corpus of language produced by the members of that culture, we might expect to see the stigmatized identity used more interchangeably with words like “gross”, “loser”, and “outcast” while the valorized identity will find itself with semantic neighbors such as “hero”, “success”, and “popular”. Note that while term frequency analysis detects *whether* terms are used, the promise of semantic similarity analysis is to understand characteristic differences in *how* terms are used.

The most common analysis of this type to date compares these measures of semantic relatedness to non-semantic measures from a different source. For instance, Caliskan et al (2017) show that associations between linguistic markers of demographic categories (e.g., uniquely Black names) and attribute words (e.g., “pleasant” and “unpleasant” words) track human implicit bias measured via IAT scores collected through Project Implicit, arguing that implicit biases seep their way into natural language. In one of their analyses, Kozlowski et al (2019) compare the semantic associations between linguistic markers of race and class with music genre names to survey-based ratings of these same associations. Van Loon and Freese (forthcoming) show that semantic associations track semantic differential scale ratings of concepts used in affect control theory, which they argue provides proof-of-concept validation for the assumption that the meanings of symbols are learned through observing and participating in symbolic exchange.

Researchers also use word embeddings to uncover dimensions of meaning we might not be able to otherwise study quantitatively. Nelson (2021), for instance, inductively analyzes associations between intersectional identities and different institutions (e.g., the domestic sphere, politics, the economy) in first-person narratives originating from the nineteenth-century American South. She finds that in this cultural context, women (and especially Black women) are more associated with the domestic sphere than men and that White folks are more associated with the cultural domain than Black folks. While the unique reach to populations and attitudes that are otherwise difficult to measure is a considerable benefit of this approach, this same feature also means that measures are difficult to validate. Nelson (2021) combats this by leveraging term frequency analyses and presenting illustrative examples from the corpus.

5.2 Between-Corpus Approach

An intuitive extension of the analyses described above would be to compare the same distance across embeddings trained on corpora produced by at different times or by different authors. If word embeddings trained on the language of members of one culture encode the meanings they share, then word embeddings trained on several corpora might reveal cultural differences in the populations that produced them (Thompson et al 2020). While the within-corpus approach described above analyzes similarity at the term level, the between-corpus uses word embeddings to obtain document-level measures that are then correlated with metadata (where the documents are typically very large and might, in other contexts, be considered an entire corpus).

For instance, while Caliskan et al (2017) were interested in the presence of bias in word embeddings, later research has examined when and where those biases exist—shedding light on its causes and consequences. Garg et al (2018) compare levels of gender and ethnic bias in newspaper articles over a century, showing that gendered meanings shifted along with the women’s movement and that racialized meanings changed along with immigration. Lewis and Lupyan (2020) show that levels of semantic gender bias across television subtitles and Wikipedia entries in different languages track gendered attitudes across thirty-nine different countries. Interestingly, they show that this semantic bias is predicted by the degree to which explicit gendering of nouns is prominent in the language (e.g. "teacher" is for all genders; "waiter" and "waitress" are explicitly gendered), arguing that these structural linguistic features drive these attitudes. Charlesworth et al (2021) are actually interested in the *lack* of variation in bias—they show that gender bias measured via word embeddings exists across language produced in different settings, at different times, and even by different age groups (children and adults).

Analyses have gone beyond associations made with demographic groups to measure how groups understand contentious concepts in general. For instance, Rudolph et al (2017) show how senators belonging to different parties use politicized terms differently in senate speeches. Studying variation in social distancing behavior across communities in the US during the early days of the 2020 Coronavirus Pandemic, van Loon et al (2020) measure county-level semantic associations between the virus and ideas of fraudulence, the political left, and benign illness in social media discourse. They argue that during this time the virus underwent a highly contingent process of social construction, that the semantic associations they measured tapped into variation in how that process unfolded across the US, and that this process had dire consequences for these communities.

At what level of granularity can one measure variation in semantic associations? Training a word embedding algorithm takes a huge amount of data, and typically training them *de novo* at the sub-national level is impractical. However, there are some approaches that can reduce the data burden associated with these models. Van Loon et al (2020), for instance, develop a straight-forward approach that shares information across sub-populations, measuring variation in semantic associations with as few as one thousand (topically relevant) tweets per county. The Bayesian approach developed by Hurtado Bodell et al (2019) allows researchers to prime models to look for variation of interest (e.g. semantic relatedness to "man" and "woman"), and it is especially helpful when analyzing small corpora. Some models such as BERT are made to be "fine-tuned" to new data (see also Dingwall and Potts 2018), though this raises some analytical concerns. With such tools it might be feasible to measure semantic associations with, say, individuals' keystroke data or the social media activity of particularly prolific users.

5.4 Discussion

Word embeddings are amazing tools that will surely be important for social science in the future, but researchers should be weary. They are generally "black boxes"—while they appear to learn valuable information from text, the process by which they learn that information (and therefore its validity) is opaque to us. This same opacity might lead researchers to give word embeddings altogether too much credit, with some even speculating these models approximate human cognition (Arseniev-Koehler and Foster 2020). We should remember that these models are wrong (but potentially useful) just like every other—similar takes exist of latent semantic analysis (Landauer and Dumais 1997; Landauer 2007), though they are now widely disregarded. We must remember that our models are simplifying tools we use to study reality, and not reality itself (Abbott 1988). Given that, it is important we understand *how* these models are wrong, so that we can take these deficiencies into account when gathering and evaluating evidence for our theoretical claims.

One growing concern among those using word embeddings is the impact of word frequency on a word's distributed representation. Specifically, word embedding models tend to segregate frequent and rare words in their estimated semantic spaces (Mu et al 2017). Wolfe and Caliskan (2021), for instance, show that linguistic bias against ethnic minorities' names is strongly correlated with how often the names appear in the underlying corpus. Van Loon et al (2022) show that this can cause omitted variable bias in social scientific analyses. They find significant and robust associations between anti-Black linguistic bias on social media and several non-linguistic measures of anti-Black animus, but also find each association to be spurious—with each being controlled away by the relative prevalences of Black and White names on social media. There are certainly other deficiencies to word embedding models, many of which we are currently unaware.

6. Conclusions

In this article I've presented a novel typology of methods for automated text analysis. My typology, unlike those presented in other reviews of automated text analysis, focuses on the nature of the feature(s) extracted from text and the way documents are implicitly represented. Specifically, I divide methods into three families. Term frequency analysis focuses on how often terms occur in each document, representing text as a single number (e.g., the proportion of words in a document that are equal to "sociology") or a bundle of such numbers. Document structure analysis attempts to inductively identify clusters of words that co-occur and differentiates documents by how aligned they are with these different clusters. Semantic similarity analysis attempts to quantify the meaning of each term by comparing the immediate contexts in which they are used, representing a document as a collection of these meanings.

I argued that this typology has two primary benefits. The first is that, by focusing on the features extracted from text by different methods, it encourages social scientists to think more deeply about operationalization and theory-building. If one is interested in measuring frames in a corpus of interest, the typology presented here gives researchers a systematic way to think about how different features of text might best operationalize frames in their corpus and then maps those features onto concrete tools. The second benefit is that it will help social scientists more effectively engage with future advances in computational linguistics. New developments might seem alien at first but knowing which linguistic features they measure (and how they do it differently) is a good heuristic to connect it to prior work and to know the importance of incorporating the advance into one's own social-scientific work.

Although there is too little empirical work to date to offer generalizable guidance on what theoretical constructs are best operationalized by different methods or families, there are clear practical tradeoffs in these choices that are worth mentioning. While some researchers might be prone to prefer the most complicated or computationally intensive method available (e.g., BERT), these same methods tend to be less transparent and more difficult to validate than less complex, less computationally intensive methods (e.g., LIWC). Additionally, although few works have performed rigorous empirical comparisons, more complicated methods tend to require more data to use. For instance, while high-quality word embeddings are often trained on corpora that contain tens or hundreds of billions of terms, closed-vocabulary term frequency analysis can be easily applied to documents the size of a tweet. Finally, the most complicated methods tend to also be the most recently developed, meaning that they have been explored and vetted less by fellow social scientists.

Throughout this article, I have mentioned family-specific issues within the field of automated text analysis. Many of their solutions require not technical improvements, but the development of norms and social scientific best practices. For instance, in term frequency analysis, there is little agreement in how to validate the correspondence between a theoretical construct and lexical prevalence. In document structure analysis, researchers are left to guess about what the appropriate size of documents is. Social scientists are discovering that word embedding models—a prominent tool for semantic similarity analysis—are subject to various biases, and best practices for their use in social science have yet to be published. Perhaps this state of normlessness in the field is excusable given how new the methodological advances driving it are, but it would nonetheless benefit from the kind of norm development seen in survey, experimental, and interview research.

Especially valuable, however, would be work that brings these different families into conversation with one another. For instance, how big of a corpus is needed to effectively use each of these families/methods? What theoretical constructs are best measured by each family? What confounding factors is each family uniquely susceptible to? Answering such questions requires the difficult work of analyzing corpora and addressing questions with multiple methods. This work would be incredibly valuable in furthering automated text analysis as a form of social scientific inquiry.

As noted in the introduction, text is becoming increasingly representative of, constitutive of, and an observable footprint of the social world. Harnessing this data at scale presents a unique opportunity for social science, but also presents unique challenges. Over the last fifty years, researchers have used the tools described in this article to effectively tackle these complexities and limitations to produce valuable scientific knowledge. Hopefully, over the next fifty years of quantitative social science, we will see more rigorous methodological validations, insightful theoretical analyses, and systematic comparisons of these methods along with their empirical application.

References

Abbott, A. (1988). Transcending general linear reality. *Sociological theory*, 169-186.

Aceves, P., & Evans, J. (2022). Mobilizing Conceptual Spaces: How Word Embedding Models Can Inform Measurement and Theory within Organization Science.

Alvero, A. J., Giebel, S., Gebre-Medhin, B., Antonio, A. L., Stevens, M. L., & Domingue, B. W. (2021). Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications. *Science advances*, *7*(42), eabi9031.

Arseniev-Koehler, A., & Foster, J. G. (2020). Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat. *arXiv preprint arXiv:2003.12133*.

Ash, E., Chen, D. L., & Galletta, S. (2022). Measuring Judicial Sentiment: Methods and Application to US Circuit Courts. *Economica*, *89*(354), 362-376.

Ashokkumar, A., & Pennebaker, J. W. (2021). Social media conversations reveal large psychological shifts caused by COVID-19's onset across US cities. *Science advances*, *7*(39), eabg7843.

Ashokkumar, A., & Pennebaker, J. W. (2022). Tracking group identity through natural language within groups. *PNAS nexus*, *1*(2), pgac022.

Bail, C. A., Brown, T. W., & Mann, M. (2017). Channeling hearts and minds: Advocacy organizations, cognitive-emotional currents, and public conversation. *American Sociological Review*, *82*(6), 1188-1213.

Bischof, J., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* (pp. 201-208).

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859-877.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, *18*, 147.

Bourdieu, P. (2019). *Distinction: A social critique of the judgement of taste* (pp. 499-525). Routledge.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. The Development and Psychometric Properties of LIWC-22.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313-7318.

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629-1635.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, *22*.

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218-240.

Choi, V. K., Shrestha, S., Pan, X., & Gelfand, M. J. (2022). When danger strikes: A linguistic tool for tracking America's collective response to threats. *Proceedings of the National Academy of Sciences*, *119*(4).

Cousin, B., Khan, S., & Mears, A. (2018). Theoretical and methodological pathways for research on elites. *Socio-Economic Review*, *16*(2), 225-249.

Danowski, J. A. (1993). Network analysis of message content. *Progress in communication sciences*, *12*, 198-221.

de Paulo Faleiros, T., & de Andrade Lopes, A. (2016). On the equivalence between algorithms for Non-negative Matrix Factorization and Latent Dirichlet Allocation. In *ESANN*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dingwall, N., & Potts, C. (2018). Mittens: an extension of glove for learning domain-specialized representations. *arXiv preprint arXiv:1803.09901*.

Dreyfus, H., & Haugeland, J. (1974). The computer as a mistaken model of the mind. In *Philosophy of psychology* (pp. 247-258). Palgrave Macmillan, London.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2018). How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., ... & Ungar, L. H. (2021). Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, *26*(4), 398.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... & Seligman, M. E. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, *26*(2), 159-169.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., ... & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, *115*(44), 11203-11208.

Elias, N. (1969). The civilizing process: Sociogenetic and psychogenetic investigations.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Fligstein, N., Stuart Brundage, J., & Schultz, M. (2017). Seeing like the Fed: Culture, cognition, and framing in the failure to anticipate the financial crisis of 2008. *American Sociological Review*, *82*(5), 879-909.

Freud, S. (1989). *The psychopathology of everyday life*. WW Norton & Company.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, *101*(1), 109.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American statistical Association*, *95*(452), 1300-1304.

Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American scientist*, *102*(6), 460.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535-74.

Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science advances*, *4*(7), eaaq1360.

Girolami, M., & Kabán, A. (2003, July). On an equivalence between PLSI and LDA. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 433-434).

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Routledge.

Goldberg, A., Srivastava, S. B., Manian, V. G., Monroe, W., & Potts, C. (2016). Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. *American Sociological Review*, *81*(6), 1190-1222.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Griffiths, Thomas L., and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences* 101.suppl 1 (2004): 5228-5235.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

Hackett, E. J., Leahey, E., Parker, J. N., Rafols, I., Hampton, S. E., Corte, U., ... & Vision, T. J. (2021). Do synthesis centers synthesize? A semantic analysis of topical diversity in research. *Research policy*, *50*(1), 104069.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.

Hart, R. P. (1984). Verbal style and the presidency: A computer-based analysis. Academic Press.

Heiberger, R. H., Munoz-Najar Galvez, S., & McFarland, D. A. (2021). Facets of Specialization and Its Relation to Career Success: An Analysis of US Sociology, 1980 to 2015. *American Sociological Review*, *86*(6), 1164-1192.

Hoffman, M. A. (2019). The materiality of ideology: cultural consumption and political thought after the American Revolution. *American Journal of Sociology*, *125*(1), 1-62.

Hoffman, M. A., Cointet, J. P., Brandt, P., Key, N., & Bearman, P. (2018). The (Protestant) Bible, the (printed) sermon, and the word (s): The semantic structure of the Conformist and Dissenting Bible, 1660–1780. *Poetics*, *68*, 89-103.

Hofmann, T. (2013). Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*.

Houston, R. (1983). Literacy and society in the West, 1500–1850. *Social history*, *8*(3), 269-293.

Hurtado Bodell, M., Arvidsson, M., & Magnusson, M. (2019). Interpretable word embeddings via informative priors. *arXiv preprint arXiv:1909.01459*.

International Telecommunications Union. (n.d.) Statistics. ITU. Retrieved September 13, 2022, from https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, *117*(19), 10165-10171.

Kim, I. S. (2017). Political cleavages within industry: Firm-level lobbying for trade liberalization. *American Political Science Review*, *111*(1), 1-20.

Kim, S. E. (2018). Media bias against foreign firms as a veiled trade barrier: Evidence from Chinese newspapers. *American Political Science Review*, *112*(4), 954-970.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, *9*(2), 137-163.

Kjell, O. N., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific reports*, *12*(1), 1-9.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905-949.

Krosnick, J. A. (1999). Survey research. *Annual review of psychology*, *50*(1), 537-567.

Landauer, T. K. (2007). LSA as a theory of meaning. *Handbook of latent semantic analysis*, *3*, 32.

Landauer, T. 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), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196). PMLR.

Lebart, L., Salem, A., & Berry, L. (1997). *Exploring textual data* (Vol. 4). Springer Science & Business Media.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788-791.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, *4*(10), 1021-1028.

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011, December). Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th international conference on data mining workshops* (pp. 81-88). IEEE.

Lynn, C. W., & Bassett, D. S. (2020). How humans learn and represent networks. *Proceedings of the National Academy of Sciences*, *117*(47), 29407-29415.

Mcauliffe, J., & Blei, D. (2007). Supervised topic models. *Advances in neural information processing systems*, *20*.

McFarland, D. A., Jurafsky, D., & Rawlings, C. (2013). Making the connection: Social bonding in courtship situations. *American journal of sociology*, *118*(6), 1596-1649.

Melamed, D. (2014). Community structures in bipartite networks: A dual-projection approach. *PloS one*, *9*(5), e97823.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics*, *41*(6), 545-569.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372-403.

Mu, J., Bhat, S., & Viswanath, P. (2017). All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3-42.

Nelson, L. K. (2021). Cycles of conflict, a century of continuity: The impact of persistent place-based political logics on social movement strategy. *American Journal of Sociology*, *127*(1), 1-59.

Nelson, L. K. (2021). Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century US South. *Poetics*, *88*, 101539.

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, *51*(1), 178-196.

Osgood, C. E. (1971). Exploration in semantic space: A personal diary 1. *Journal of Social Issues*, *27*(4), 5-64.

Pauca, V. P., Shahnaz, F., Berry, M. W., & Plemmons, R. J. (2004). Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM international conference on data mining* (pp. 452-456). Society for Industrial and Applied Mathematics.

Paxton, P., Velasco, K., & Ressler, R. W. (2020). Does use of emotion increase donations and volunteers for nonprofits?. *American Sociological Review*, *85*(6), 1051-1083.

Payne, S. L. B. (2014). *The art of asking questions*. Princeton University Press.

Pennebaker, J. W. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour research and therapy*, *31*(6), 539-548.

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition & Emotion*, *10*(6), 601-626.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC [Computer software]. Austin, TX: LIWC.net.

Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of personality and social psychology*, *72*(4), 863.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin

Pennington, J., Socher, R., & Manning, C. D. (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).

Pew Research Center. (2008, September). "Internet and Email Use for Work." Pew Research Center, Washington, D.C. https://www.pewresearch.org/internet/2008/09/24/internet-and-email-use-for-work/#fn-613-14.

Poe, M. T. (2010). *A History of Communications: Media and Society from the Evolution of Speech to the Internet*. Cambridge University Press.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988-1003.

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013, December). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation* (Vol. 4, pp. 1-20).

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, *58*(4), 1064-1082.

Rocklage, M. D., Rucker, D. D., & Nordgren, L. F. (2021). Mass-scale emotionality reveals human behaviour and marketplace success. *Nature human behaviour*, *5*(10), 1323-1329.

Rosenfeld, M. J., Thomas, R. J., & Hausen, S. (2019). Disintermediating your friends: How online dating in the United States displaces other ways of meeting. *Proceedings of the National Academy of Sciences*, *116*(36), 17753-17758.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2012). The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*.

Roser, M., & Ortiz-Ospina, E. (2016). Literacy. *Our World in Data.*

Rudolph, M., Ruiz, F., Athey, S., & Blei, D. (2017). Structured embedding models for grouped data. *Advances in neural information processing systems*, *30*.

Rule, A., Cointet, J. P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, *112*(35), 10837-10844.

Ryle, G. (2009). *Collected Essays 1929-1968: Collected Papers Volume 2*. Routledge.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual review of sociology*, *29*(1), 65-88.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712.

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. *Political communication*, *35*(1), 50-74.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, *63*(4), 517-522.

Taddy, M. (2012, March). On estimation and selection for topic models. In *Artificial Intelligence and Statistics* (pp. 1184-1193). PMLR.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467-484.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

Tavory, I., & Timmermans, S. (2014). *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, *4*(10), 1029-1038.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information processing & management*, *50*(1), 104-112.

van Loon, A., & Freese, J. (Forthcoming). Word Embeddings Reveal How Fundamental Sentiments Structure Natural Language. *American Behavioral Scientist*.

van Loon, A., Giorgi, S., Willer, R., & Eichstaedt, J. (2022). Regional Negative Bias in Word Embeddings Predicts Racial Animus--but only via Name Frequency. *arXiv preprint arXiv:2201.08451*.

van Loon, A., Stewart, S., Waldon, B., Lakshmikanth, S. K., Shah, I., Guntuku, S. C., ... & Eichstaedt, J. (2020). Explaining the Trump Gap in Social Distancing Using COVID Discourse. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020***.**

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009, June). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105-1112).

Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Wolfe, R., & Caliskan, A. (2021). Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. *arXiv preprint arXiv:2110.00672*.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445-1456).