UNIVERSITY OF CALIFORNIA

Los Angeles

Measuring and Modeling Cultural Meaning in Language

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Sociology

by

Alina Ashelman

2022

ABSTRACT OF THE DISSERTATION

Measuring and Modeling Cultural Meaning in Language

by

Alina Ashelman
Doctor of Philosophy in Sociology
University of California, Los Angeles, 2022
Professor Jacob Gates Foster, Chair

Measuring meaning is a longstanding methodological problem in social science – especially meaning in text data. It is also a theoretical problem: measurement requires us to specify (Merton 1948) the theoretically vague concept of meaning itself. Cultural sociologists have spent decades trying to clarify aspects of meaning, such as the extent to which meaning is structured and stable across contexts and the extent to which meaning is stored in our minds versus in external symbols (e.g., images, words, and even parts of words such as suffixes). Meanwhile, recent advances in computer science offer new measures and formal models of meaning in text data. For example, word embeddings quantitatively model the meaning of words in text data. This dissertation capitalizes on such recent advances in computer science to contribute to theoretical and methodological work on meaning in cultural sociology. The first paper theorizes the kinds of sociological meaning that word embeddings operationalize and describes how cultural sociologists can use word embeddings to empirically investigate meaning in text. The second paper uses word embeddings to empirically investigate the extent to which syntactically gendered language (e.g., “police*man*”) conveys gendered semantic information. Finally, paper three develops a novel computational approach to measure latent thematic meaning in large-scale text data, integrating word embedding and topic modeling approaches to measuring meaning in text data. The third paper then applies this new approach to identify latent themes in a large, underutilized source of text data on violent death in the U.S.

This dissertation of Alina Ashelman is approved.

Jessica Lyn Collett

Morteza Dehghani

Omar Alcides Lizardo

Jacob Gates Foster, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

List of Tables and Figures vii

Acknowledgments ix

Chapter 1. Introduction 1

Chapter 2. Theoretical foundations and limits of word embeddings: What types of meaning can they capture? 8

ABSTRACT 8

A PRIMER ON WORD EMBEDDINGS 11

Word Embeddings as Social Science Methods 13

Approaches to Compute Word Embeddings Commonly used in Social Science 16

A PRIMER ON STRUCTURAL LINGUISTICS 19

OPERATIONALIZING STRUCTURAL LINGUISTICS WITH WORD EMBEDDINGS 22

Modeling Language as a Relational System with Word Embeddings 22

Modeling Language as a Coherent System with Word Embeddings 25

Modeling Language as a Static System with Word Embeddings 26

FOUR CRITIQUES OF STRUCTURAL LINGUISTICS 27

Critiques of Purely Relational Approaches to Meaning 28

Critiques of Binary Oppositions 31

Critiques of Coherence 34

Critiques of a Static Lens on Language 39

DISCUSSION 42

Directions for Future Social Science Research using Embeddings 44

Conclusions 46

Chapter 3. Are both policemen and policewomen police officers? The relationship between meaning and form in gendered language 48

ABSTRACT 48

BACKGROUND 52

Cultural Symbols and Symbolic Systems 52

Gendered Language 57

METHODS 67

Approach 67

News Text Data 67

Identifying Occupational Titles and Manually Annotating Titles for Gender Suffixes.................68
Modeling Gendered Linguistic Meanings of Occupational Titles with Word Embeddings...........70
Statistical Analyses...77
RESULTS...78
Occupational Titles' Gendered Meanings correspond to their Gender Suffixes............................79
Occupational Titles with Suffixes Carry More Gendered Meanings than Titles without Gender Suffixes...80
Even Occupational Titles without Suffixes Carry Gendered Meanings...82
These Results Hold even if Occupational Titles are Processed as Compound Words....................83
DISCUSSION...86
Limitations and Future Research...89
Implications...91
Chapter 4. Integrating topic modeling and word embedding to characterize violent deaths...........93
ABSTRACT...93
INTEGRATING TOPIC MODELING AND WORD EMBEDDING WITH THE DISCOURSE ATOM TOPIC MODEL (DATM)...97
Identifying Topics in Semantic Space...98
Moving from Semantic Space to Text Data, and Back...99
TOPICS IN DESCRIPTIONS OF VIOLENT DEATH...104
Topics and Latent Semantic Dimensions...107
DISCUSSION...114
SUPPLEMENTARY INFORMATION...115
Data...115
Training the Word Embedding...118
Connections between Continuous-Bag-of-Words (CBOW) Word Embeddings and DATM........120
Identifying DATM Topics in a Word Embedding with K-SVD...121
DATM Model Quality and Selecting the Number of Topics...128
Comparing DATM to Other Topic Models...149
LDA Topic Modeling on NVDRS Narratives...151
Semantic Dimensions Beyond Gender...158
Chapter 5. Conclusion...160
WORKS CITED...163

List of Tables and Figures

Tables

Table 1. Gender Meanings of 460 Occupational Titles 80

Table 2. Overlaps and differences among the top 50 closest words to: Police officer, Policewoman, and Policeman........ 85

Table 3. Sample Topics within Narratives of Violent Death. 107

Table 4. Characteristics of Violent Deaths with Two Selected Topics........ 113

Table 5. Characteristics of Sample of Violent Deaths, drawn from the National Violent Death Reporting System........ 118

Table 6. All 225 Topics identified using the Discourse Atom Topic Model 133

Table 7. All 100 Topics identified using LDA Topic Modeling........ 153

Table 8. Random Sample of 15 Topics identified using LDA Topic Modeling with 225 Topics 158

Figures

Figure 1. Conceptual illustration of a latent dimension in semantic space corresponding to gender (left) and morality (right). 13

Figure 2. Gendered meanings for a random sample of 50 occupational titles, where meaning based on use in the New York Times 2002-2007. The gendered meaning of an occupational title is standardized based on the gendered meanings of all words in the corpus. 79

Figure 3. Gendered meanings of 460 occupational titles with and without gender suffixes, using FastText-based measures of gendered meaning trained on New York Times data. Gendered meaning (y-axis) is the number of standard deviations more feminine (+) or more masculine (-) compared to the average word in the corpus. 82

Figure 4. Gendered meanings of 460 occupational titles with and without gender suffixes, using word2vec based measures of gendered meaning trained on New York Times data. Gendered meaning (y-axis) is the number of standard deviations more feminine (+) or more masculine (-) compared to the average word in the corpus. 84

Figure 5. Prevalence of 225 Topics in Narratives of 272,964 Decedents of Violent Death, by Manner of Death. Note: Each row represents the fraction of narratives with a given topic by manner of death, row standardized across all manners of death. 106

Figure 6. Latent Gendered Meanings of Topics vs Prevalence of Topics in Female vs Male Decedents' Narratives. Notes: N= 225 topics. For clarity, labels are shown only for topics with high or low gender meanings or gender prevalence ratios; overlapping labels are removed. The y axis represents cosine similarity between a given topic and the gender dimension in semantic space. The x axis represents the ratio of female decedents' narratives containing a given topic compared to narratives of male decedents. 110

Figure 7. Decomposing the Embedding Matrix into a Dictionary of Topics and Coefficients. 124

Figure 8. Figure A-D. Measures of Model Quality (Coherence, Diversity, Coverage, and RMSE) against the Number of Topics in the Discourse Atom Topic Model. 132

Acknowledgments

I am grateful to my committee, Jessica Collett, Omar Lizardo, Morteza Dehghani, and Jacob Foster, for their mentorship and support throughout the doctoral program, and for so profoundly inspiring me with their scholarship and teaching. I especially thank Jacob Foster, chair of the committee and my primary advisor, for believing in me and for making graduate school a transformative and fulfilling intellectual adventure. I also thank the many scholars who mentored me across college and graduate school, including Andrei Boutyline, Rachel Best, Susan Cochran, Vickie Mays, Kai-Wei Chang, Abigail Saguy, Megan Moreno, Tyler McCormick, and Hedwig Lee. I am deeply grateful to my wonderful family for their endless support.

This dissertation work was funded by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650604 and a Dissertation Year Fellowship from UCLA Graduate Division. It also won the 2021 Best Dissertation-in-Progress Award from the American Sociological Association, Section on Mathematical Sociology.

Previous versions of the three papers in this dissertation were presented at various conferences. Versions of paper two were presented at the International Conference on Computational Social Science in 2019; the Text Analysis across Domains Conference at the University of California, Berkeley in 2019; and the University of Washington Gender Working Group in 2019. Versions of paper three were presented at the American Sociological Association Meeting, Session on Computational Sociology in 2020, and at the University of California, Berkeley Computational Text Analysis Working Group in 2020. I thank attendees at all these presentations for their feedback and enthusiasm. I thank everyone in the course Social Psychology of Gender for their encouraging feedback on an early version of paper two.

Paper one has a corresponding preprint on *SocArXiv* (https://osf.io/preprints/socarxiv/vrwk3/). The DOI is 10.31235/osf.io/vrwk3. It is under a third round of peer review for publication at *Sociological Methods & Research.* I am very grateful to five anonymous reviewers at this journal for their very extensive and very constructive feedback on paper one, which substantially improved the paper.

Paper three was published as an article titled “Integrating Topic Modeling and Word Embedding to Characterize Violent Deaths” in the *Proceedings of the National Academy of Sciences (PNAS)* on March 3, 2022 (volume 119, number 10, page 2108801119). The DOI is https://doi.org/10.1073/pnas.2108801119. A preprint is also posted on *ArXiv* (https://arxiv.org/abs/2106.14365) and *SocArXiv* (https://osf.io/preprints/socarxiv/nkyaq/). This paper is first-authored; Susan Cochran, Vickie M. Mays, Kai-Wei Chang, and Jacob G. Foster are coauthors (in that order). All authors contributed to the design of the research and interpretation of results. Vickie M. Mays and Susan Cochran obtained access to the data. Susan Cochran and I cleaned the data. Jacob Foster and I conceptualized the core method. I implemented analyses, wrote the code, and wrote the first draft. All authors edited the paper. I thank the anonymous reviewers at *PNAS* and Sanjeev Arora for their detailed and constructive feedback on the paper.

Alina Ashelman (previously Arseniev-Koehler)
Vita

EDUCATION

2017	M.A., Sociology Committee: Jacob Foster (chair), Abigail Saguy	University of California, Los Angeles
2014	B.A. with Honors, Sociology	University of Washington

FELLOWSHIPS AND AWARDS (SELECTED)

2021-2022	Dissertation Year Fellowship – UCLA Graduate Division
2021	Best Dissertation-in-Progress Award – American Sociological Association (ASA), Section on Mathematical Sociology
2020	Outstanding Graduate Student Paper Award – American Sociological Association (ASA), Section on Mathematical Sociology
2017-2020	Graduate Research Fellowship – National Science Foundation

PUBLICATIONS (SELECTED)

Seamans, Marissa, Vickie Mays, **Alina Arseniev-Koehler**, and Susan Cochran. "Prevalence of prescription and illicit substances in the environment among suicides of non-poisoning means in the National Violent Death Reporting System 2003-2017." Forthcoming in *The American Journal of Drug and Alcohol Abuse*.

Arseniev-Koehler, Alina, Susan Cochran, Vickie Mays, Kai-Wei Chang, and Jacob Foster. 2022. "Integrating topic modeling and word embedding to characterize violent deaths." *Proceedings of the National Academy of Sciences*. 119(10): 2108801119.

Arseniev-Koehler, Alina, Jacob Foster, Vickie Mays, Kai-Wei Chang, and Susan Cochran. 2021. "Aggression, escalation, and other latent themes in legal intervention deaths of non-Hispanic Black and White men: Results from the 2003-2017 NVDRS." *American Journal of Public Health*. 111: S107-S115.

Arseniev-Koehler, Alina and Jacob Foster. 2021. "Sociolinguistic Properties of Word Embeddings." Pp 464-477 in *The Atlas of Language Analysis in Psychology*, edited by Morteza Dehghani and Ryan L. Boyd. Guilford Press

Boutyline, Andrei, Devin Cornell, and **Alina Arseniev-Koehler**. 2021. "All Roads Lead to Polenta: Cultural Attractors at the Junction of Public and Personal Culture." Special issue on culture and cognition in *Sociological Forum*. 36 (S1): 1419-1445.

Ankith Uppunda, Susan Cochran, Jacob Foster, **Alina Arseniev-Koehler**, Vickie Mays, and Kai-Wei Chang. 2021. "Adapting Coreference Resolution for Processing Violent Death Narratives." Pp 4553-4559 in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Larimore, Savannah, Ian Kennedy, Breon Haskett, and **Alina Arseniev-Koehler**. "Reconsidering Annotator Disagreement about Racist Language: Noise or Signal?" 2021. Pp 81-

90 in *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*.

Moreno, Megan A, Molly Adrian, Megan A Wilt, Adrienne Ton, **Alina Arseniev-Koehler**, Elizabeth McCauley, and Ann Vander Stoep. 2021. "Displayed Depression Symptoms on Facebook at Two Time Points: Content Analysis." *JMIR Formative Research*.5(5): e20179.

Arseniev-Koehler, Alina, Hedwig Lee, Tyler McCormick, and Megan Moreno. 2014. "#Pro-Ana: Pro-Eating Disorder Socialization on Twitter." *Journal of Adolescent Health*. 58(6):659-664.

WORK UNDER REVIEW

Arseniev-Koehler, Alina. "Theoretical foundations and limits of word embeddings: What types of cultural meaning can they capture?" R&R at *Sociological Methods and Research*.

Arseniev-Koehler, Alina and Jacob Foster. "Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat." R&R at *Sociological Methods and Research.*

Boutyline, Andrei, Alina Arseniev-Koehler, and Devin Cornell. "School, Studying, and Smarts: Gender Stereotypes and Education Across 80 Years of American Print Media, 1930-2009." R&R at *Social Forces.*

Best, Rachel and Alina Arseniev-Koehler. "Stigma's uneven decline." R&R at *American Sociological Review.*

CONFERENCE PRESENTATIONS (SELECTED)

"Theoretical foundations and limits for word embeddings: What types of cultural meaning can they capture?" 2021. American Sociological Association Meeting. Virtual due to Covid-19.

"Discourses of Death: Introducing a new method to model topics in word embeddings." 2020. American Sociological Association Meeting. Virtual due to Covid-19.

"The Gendered Meanings of Gender-Fair Language." 2019. Text Analysis across Domains (TextXD) Conference. UC Berkeley, CA.

"Gender equal or gender neutral? Gender connotations in occupational titles." International Conference on Computational Social Science (IC2S2). 2019. Amsterdam, Netherlands.

POSITIONS (SELECTED)

Research Associate I. University of Michigan, Ann Arbor. 6/2019 – 6/2022.

Organizer and Instructor. Summer Institute in Computational Social Science (SICSS) Los Angeles. Annually in Summer 2019, 2020, and 2021.

Visiting Research Assistant. University of Southern California Institute for Creative Technologies. 6/2017 – 9/2017.

Clinical Research Assistant. Seattle Children's Hospital Research Institute. 7/2014 – 8/2015.

Chapter 1. Introduction

This dissertation is about one of the most fundamental, and elusive, concepts in sociology: Meaning. The meanings attached to symbols – such as words, icons, and suffixes – guide how we think, interact, share information, and act. Meanings also shape our well-being and life chances and do so in ways that contribute to inequality. Salient examples include the stereotypes and implicit associations around peoples' bodies' colors, weights, shapes, parts, and textures. To the extent that meanings are shared across individuals and participate in social life, they are also cultural (Foster 2018; Sperber 1985; Strauss and Quinn 1997:6–7, 85). While there is little doubt about the importance of meaning in sociology, it remains unclear exactly what it is and how it operates in our minds and through symbols and symbolic systems (e.g., language). There are also many remaining questions as to how to empirically measure meaning and meaning-making – especially, in the context of written language (Mohr 1998; Mohr et al. 2020; Mohr and Ghaziani 2014).

Meanwhile, recent advances in computer science offer new approaches to quantitatively capturing meaning in text data. Most prominently, word embedding (e.g., Devlin et al. 2019; Mikolov, Sutskever, et al. 2013; Pennington, Socher, and Manning 2014) and topic modeling (e.g., Blei 2012) quantitatively represent the meaning of words, themes, semantic dimensions, and other semantic structures. They also learn and process these representations from interactions with natural language, such as from predicting missing words while "reading" a corpus. Many of these methods were developed in computer science to accomplish downstream tasks using text data, such as machine translation, recommendation searches, sentiment analysis, and text summarization. However, their success on these downstream tasks attests to their effectiveness at quantitatively representing meaning in text data in human-like ways (e.g., Caliskan, Bryson, and Narayanan 2017; Tomas Mikolov, Chen, et al. 2013). As this dissertation shows, their successes also offer sociologists new empirical opportunities to capture meaning in text data, and new theoretical opportunities to clarify what meaning is and how it works.

Measures and models of meaning in text data, such as word embedding and topic modeling, offer theoretical opportunities to specify and clarify vague notions of meaning itself (Merton 1948). For example, topic models offer a formal model for the theoretical construct of "frames" in cultural sociology (DiMaggio, Nag, and Blei 2013). More recently, neural word embeddings (those trained with artificial neural networks) offer a formal, mathematical model for how individuals learn and process schema from their cultural environment (Arseniev-Koehler and Foster 2020).

The first paper in this dissertation contributes to such efforts to capitalize on the ties between word embedding methods and sociological theorizing of meaning. In this theory paper, I argue that word embeddings operationalize meaning in a structuralist fashion (a conception of meaning that was heavily critiqued in cultural sociology).[1] Then, I argue how embeddings (1) succumb to certain critiques of structuralism and (2) provide new solutions to other critiques of structuralism. More broadly, this paper argues that word embeddings enable us to reimagine the

[1] By structuralism, I refer to semiotic (or French) structuralism, most exemplified in sociology by the work of Lévi-Strauss. This is distinguished from British Structuralism, which was more concerned with social networks and social network analysis. In addition, while I focus on the Saussurean influences on French Structuralism, there were many other influences as well (Maryanski and Turner 1992; Stoltz 2019).

study of meaning, just as social network analysis reimagined the study of social ties (Borgatti et al. 2009). Paper one is currently under a third round of peer review at a high-impact journal. In a book chapter with Jacob Foster in *Language Analysis in Psychology*, edited by Morteza Dehghani and Ryan Boyd, we expanded on several points raised in this paper.

Like other papers in this dissertation, Paper 1 explicitly focuses on the case of meaning as it pertains to language. Language is arguably the most well-studied and salient symbolic system. Many core sociological perspectives on meaning also stem from the study of language (for a historical account of structuralism see Dosse 1997; Joas and Knobl 2009). Similarly, as I highlight at various points in Paper one, conclusions drawn from studying words and language can be extended to other symbols and symbolic systems, such as images, gestures, and sound.

New methods to measure meaning in text data also offer new empirical opportunities for social scientists to study how meaning works (e.g., Best and Arseniev-Koehler 2022; Boutyline, Arseniev-Koehler, and Cornell 2020; Chakrabarti and Frye 2017; Dehghani et al. 2017; DiMaggio et al. 2013; Jones et al. 2020; Kozlowski, Taddy, and Evans 2019). For example, in a co-authored paper with Rachel Best (2022), we use word embedding methods to investigate how and why stigmatizing meanings of diseases change across time. We operationalize four stigmatizing meanings identified by prior work, such as the extent to which a disease signals moral failing or repulsion, and we measure the extent to which 107 diseases evoke these meanings across four decades (1980-2016) of U.S news (including 27 news outlets). Our findings reveal a changing landscape of disease stigma: Most diseases carried potent stigmatizing meanings in the 1980s, and most diseases shed these meanings across time. However, the stigmatizing meanings of several groups of diseases, including mental illnesses, remained strikingly stable. These findings illuminate conflicting results in prior work on disease stigma across time, which often focused on a handful of diseases or time points. In another co-authored paper with Andrei Boutyline, we similarly leverage word embeddings to investigate change and stability in gender stereotypes related to education, across nearly a century of news, magazines, fiction, and media (Boutyline et al. 2020). As these studies illustrate, methods such as word embeddings enable social scientists to investigate empirical questions about cultural meaning which may be impractical or impossible with traditional approaches such as surveys. They also allow for a new vantage point from which to capture meaning in the cultural environment (Bail 2014).

In Paper two, I use word embeddings to investigate the relationship between words' gendered forms (e.g., "police*man*," "congress*woman*," "actr*ess*") and their gendered meanings. An abundance of work in psychology and cognitive linguistics illustrates that gendered form is associated with individuals' meanings of these words (Boroditsky, Schmidt, and Phillips 2003). Such work often assumes that this relationship is driven by the patterned ways in which gendered forms are used (Atagi, Sethuraman, and Smith 2009) – that people learn gendered meanings of words from the patterned ways that words are used in natural language. However, it has not yet been explicitly tested whether gendered form corresponds to how words are used in natural language (i.e., linguistic meaning). I use word embeddings to empirically examine this relationship, capitalizing on the fact that neural word embeddings can *only* capture the linguistic meaning of words and do not have extra-linguistic experiences (unlike humans). I focus on natural language data from news and media from a variety of sources, and I focus on occupational titles with and without gender suffixes. My results validate assumptions from prior work, demonstrating that occupational titles' masculine/feminine gendered forms correspond to their gendered linguistic meaning. In addition, I find that the relationship between grammatical gender and gendered linguistic meaning holds in both neural word embeddings that account for form (i.e., they are exposed to suffixes as pieces of words), and neural word embeddings that do *not* (i.e., that only see compound words). This suggests that gendered meanings are not driven by way *suffixes* are used, per se, but the ways that the entire words with (vs without) suffixes are used. For cultural sociology, gendered language offers a case to reconsider the prevailing notion that symbols' forms (e.g., their letters or sounds) are unrelated to their meanings. My empirical results also contribute to scholarly debates around how to minimize gender biases in language. More generally, gendered language presents an important and well-documented case to consider the relationship between the meanings of symbols in cultural objects, like news, and individuals' internalized meanings.

Despite the substantial methodological progress in measuring meaning in text data across the disciplines, current approaches have numerous limitations to be addressed. For example, it is notoriously challenging to tune topic models to produce human-interpretable topics (Boyd-Graber, Mimno, and Newman 2014). Further, these methods are often disconnected from one another. For instance, while embeddings offer rich, nuanced representations of semantic information in words and sentences, they do not explicitly capture latent topics. While topic models enable analysts to connect words, latent topics, and documents in a corpus, they do not capitalize on the successes of word embeddings at representing semantic information. These methodological limitations offer opportunities for innovation and are the focus of Paper three.

Paper three contributes to ongoing work in social science and computer science to address limitations of existing measures in meaning in text data. This paper was recently published in the high impact journal *PNAS* (Arseniev-Koehler et al. 2022). This methodological paper introduces the Discourse Atom Topic Model: A robust and flexible new computational approach to jointly model the latent meaning of words, documents, *and* topics in large-scale text data. The model is useful because it integrates two core (but separate) text analysis methods (topic modeling and word embedding) into a cohesive approach. To do so, it imports a generative model from machine-learning theory (Arora, Li, Liang, et al. 2016; Arora, Liang, and Ma 2017). This generative model describes the probability a word is produced, given the "gist" or latent topic being talked about. It operationalizes "gist" using techniques to embed sentences into semantic space, which are motivated by the same generative model (Arora et al. 2017; Ethayarajh 2018). The model produces distinctly interpretable topics, and it enables researchers to distill documents down to a sequence of topics (or ignoring the order, a distribution of topics). Further, with this model, we can generalize analytic techniques typically applied to *words* in word embeddings – such as identifying latent gendered meanings of words – to study *topics*.

To illustrate the utility of the Discourse Atom Topic Model, in Paper three we then use it to identify themes and gendered patterns in a large, underutilized source of text data on violent death in the U.S. These descriptions are collected as part of violent death surveillance by the Centers for Disease Control and Prevention (CDC). As for many administrative datasets, text variables in this surveillance data remain underutilized due to methodological barriers. Our results illustrate latent topics in the narratives, many of which are not yet part of the standard menu of characteristics used to classify violent deaths in the CDC. Like many other characteristics of violent death, these topics are also distinctly patterned by gender. These gender patterns reflect both differences in how men and women die from violent death and differences in how men's and women's violent deaths are described by public health workers at the CDC. In another article with the same team, we used this topic model to investigate racial/ethnic differences in police shootings and other deaths resulting from legal intervention (Arseniev-

Koehler et al. 2021). We draw our data from the same CDC surveillance system. Our results contribute nuanced evidence that during these lethal encounters, Black male decedents posed little threat to legal authorities (compared to White males). For example, we found that written summaries of Black males' deaths were significantly *less* likely to mention a topic about physically aggressive actions. These findings resonate with prior work that legal authorities perceive danger differently during encounters with Black males, compared to encounters with White males. This paper was recently published in *The American Journal of Public Health*.

This dissertation is outlined as follows. Paper one theorizes the kinds of sociological meaning that word embeddings operationalize and describes how cultural sociologists can use word embeddings to empirically and theoretically study meaning. Paper two uses word embeddings to empirically investigate the relationship between gendered language (e.g., "police*man*") and gendered linguistic meanings. Finally, Paper three develops Discourse Atom Topic Modeling (DATM), and then uses it to uncover latent themes in a large, underutilized source of text data on violent death in the U.S. It also examines how these themes are loaded with meanings of gender. To conclude, I summarize the contributions of these papers and I highlight connections between sociological work on meaning and broader interdisciplinary questions about meaning.

Chapter 2. Theoretical foundations and limits of word embeddings: What types of meaning can they capture?

ABSTRACT

Measuring meaning is a central problem in cultural sociology. Word embeddings may offer powerful new tools to do so. But like any tool, they build on and exert theoretical assumptions. In this paper, I theorize the ways in which word embeddings model three core premises of a structural linguistic theory of meaning: That meaning is coherent, relational, and can be analyzed as a static system. In certain ways, word embeddings are vulnerable to the enduring critiques of these premises. In other ways, word embeddings offer novel solutions to these critiques. More broadly, formalizing the study of meaning with word embeddings offers theoretical opportunities to clarify core concepts and debates in cultural sociology, such as the coherence of meaning. Just as network analysis specified the once vague notion of social relations (Borgatti et al. 2009), formalizing meaning with embeddings can push us to specify and reimagine meaning itself.

Measuring, modeling, and understanding how meaning operates are several of the most prominent and longstanding endeavors of sociology (e.g., Mohr 1998; Mohr et al. 2020). In recent years, word embedding methods reinvigorated the study of meaning (e.g., Arseniev-Koehler and Foster 2020; Boutyline et al. 2020; Charlesworth et al. 2021; Jones et al. 2020; Kozlowski et al. 2019; Nelson 2021; Stoltz and Taylor 2019, 2020; Taylor and Stoltz 2020). These methods computationally model the semantic content of words in large-scale text data. Despite their promise, it remains unclear what kind of meaning word embeddings capture – or whether they capture any meaning at all. To employ these tools rigorously, it is paramount that we clarify what they operationalize. In this paper, I critically evaluate the possibility that word embeddings operationalize an influential theory of language: structural linguistics.

Word embedding, and in particular the word2vec word embedding algorithm, revolutionized how computers learn and process human language (Mikolov, Chen, et al. 2013; Mikolov, Sutskever, et al. 2013). Indeed, since word2vec was published in 2013 it has been cited nearly 33,000 times (Mikolov, Sutskever, et al. 2013). Rapid advances in computer science have yielded a tremendous variety of word embedding approaches, and strategies to model language more generally (e.g., Devlin et al. 2019). Meanwhile, social scientists have imported word embeddings to analyze text data at scale (e.g., Best and Arseniev-Koehler 2022; Boutyline et al. 2020; Boutyline, Cornell, and Arseniev-Koehler 2021; Charlesworth et al. 2021; Garg et al. 2018; Haber 2021; Hamilton, Leskovec, and Jurafsky 2016; Jones et al. 2020; Kozlowski et al. 2019; Linzhuo, Lingfei, and Evans 2020; Martin-Caughey 2021; Nelson 2021; Rozado 2019; Stoltz and Taylor 2019, 2020; Taylor and Stoltz 2020). For example, these methods are used to capture stereotypes encoded in media language across time, offering a historical lens into stereotypes despite the absence of corresponding survey data (e.g., Jones et al. 2020). The premises of word embeddings may also be generalized beyond language to model other cultural systems (e.g., Arronte Alvarez and Gómez-Martin 2019; Chuan, Agres, and Herremans 2020). As such work highlights, word embeddings offer exciting new lenses to study language – and perhaps other symbolic systems – across time and space.

Researchers employing word embedding methods often note their affinities to a century-old theoretical perspective on language: structural linguistics (e.g., Baroni, Dinu, and Kruszewsk 2014; Faruqui et al. 2016; Günther, Rinaldi, and Marelli 2019; Kozlowski et al. 2019). Structural linguistics envisions language as a symbolic system comprised of various linguistic units (e.g., words, phonemes, or suffixes). These units are defined by their relationships to other units in the system (rather than by their reference to physical entities in the world). For instance, a word may be characterized by its co-occurrence relationships with other words. Social scientists generalized the premises of structural linguistics to study non-linguistic symbolic systems (e.g., kinship systems, Lévi-Strauss 1963:35) and to theorize culture as a symbolic system more broadly. This intellectual movement is often referred to as semiotic structuralism, French structuralism, or just structuralism.

Here, I critically examine the affinities between the way that word embeddings model words' semantic information and the structural linguistic perspective on word meaning. I focus on how word embeddings operationalize (or do not operationalize) core premises of structural linguistics. The first major contribution of this paper is to highlight that the extent to which contemporary word embedding methods operationalize structural linguistics depends on the way these methods are used, the specific embedding algorithm used, and even an analysts' own interpretation of "meaning" in the algorithm.

Structural linguistics (and structuralist approaches to culture more broadly) are both profoundly influential and intensely critiqued across the social sciences (Dosse 1997; Sewell 2005). For instance, critics questioned the plausibility of a coherent symbolic system (e.g., Martin 2010; Swidler 1986), noting that cultural symbols (like words) are often used in strikingly incoherent ways (e.g., Swidler 2013). Given the parallels between word embeddings and structuralism, do word embeddings also model language in a way that is vulnerable to the critiques of structuralism? The second major contribution of this paper is to evaluate the ways in which word embeddings *succumb to* and *overcome* the limitations of structural linguistics.

To begin, I first briefly review background information on word embedding methods as they are used in social science. Second, I review structural linguistics, focusing on three of its core premises (that language is relational, coherent, and should be studied as a static system). Third, I critically examine the ways in which word embeddings operationalize each of these three premises. Fourth, I consider critiques of these premises and evaluate whether and how each critique applies to word embeddings. In the discussion, I highlight implications and directions for future sociological research with word embedding methods.

A PRIMER ON WORD EMBEDDINGS

Word embeddings are quantitative representations of the semantic information of words, which are computed based on how those words are used in a text dataset (i.e., training data). Examples of text datasets include a corpus of news articles, social media posts, government records, or product reviews. Word embeddings aim to represent words as vectors (i.e., arrays of *N* numbers) where words that are used in more similar contexts in this training data are assigned more similar vectors. Word vectors may also be understood geometrically — as positions in an *N*-dimensional space. The dimensionality of the space (and thus all word vectors) is pre-set by the algorithm or the researcher – often, at a few hundred dimensions (Rong 2014). The information captured by dimensions is latent and often uninterpretable to humans; they are identified by word embedding algorithm as it organizes words in space through a training algorithm. Two examples of training algorithms, word2vec (Mikolov, Sutskever, et al. 2013) and GloVe (Pennington et al. 2014), will be detailed shortly. Since vectors locate positions in space, similarity and distance are interchangeable. Words with more similar vector representations are also *closer* in space. This similarity, or distance, is commonly measured with cosine similarity. The collection of the word-vectors may also be referred to as the *trained* word embedding, a semantic space, or, as just a word embedding.

A body of research finds that the semantic information captured by well-trained word embeddings corresponds to humans' own understandings of words (for a review, see Mandera, Keuleers, and Brysbaert 2017). For example, in well-trained word embeddings, the cosine similarities between word vectors strongly correlate to human-rated similarities between words (e.g., Caliskan, Bryson, and Narayanan 2017; Pennington et al. 2014). Further, while word embeddings are trained to represent words as positions in space (i.e., as vectors), empirical work finds that the space *between* word vectors may also capture semantic information. Most famously, the direction to travel in semantic space between the word vectors "king" and "queen" is often similar to the direction to travel between "man" and "woman." That is, the difference between the locations of "king" and "queen" in semantic space is similar to the difference between "man" and "woman" (Mikolov, Sutskever, et al. 2013). The difference may be measured by subtracting the corresponding word vectors, e.g., "woman" – "man."

The result of this subtraction is a vector that may be interpreted as a latent line in space, pointing to words about women at one pole and words about men at the other pole (see Figure 1). In fact, a variety of concepts beyond gender may be encoded as latent dimensions in space (Bolukbasi et al. 2016; Kozlowski et al. 2019; Mikolov, Sutskever, et al. 2013). This property of word embeddings attests to their ability to encode semantic information in rich and nuanced ways. Furthermore, as described next, this property makes word embeddings a remarkably useful social science method.

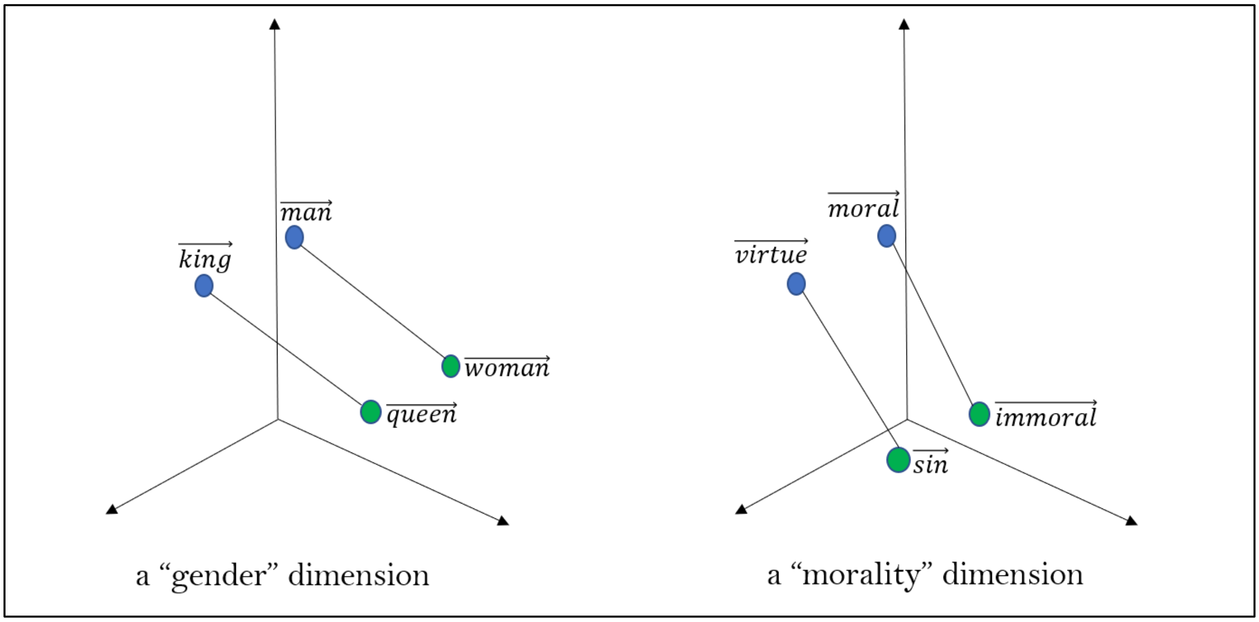

Figure 1. Conceptual illustration of a latent dimension in semantic space corresponding to gender (left) and morality (right).

Word Embeddings as Social Science Methods

In recent years, word embedding has exploded as an exciting new method in social science (for a review, see Stoltz and Taylor 2021). One popular analytic approach is to deductively identify a latent semantic dimension (e.g., gender, social class, sentiment or morality) in the embedding space, and then examine how a set of sociologically interesting keywords (e.g., occupational roles or stereotypical traits) are positioned along this dimension (e.g., Kozlowski et al. 2019). For example, researchers may identify a line corresponding to gender (e.g., by subtracting the word vector "she" from "he") and then examine how close occupational terms are to the pole about women versus the pole about men (Bolukbasi et al. 2016). Effectively, this approach enables analysts to compute the association of a word (or set of words) with a latent concept in text data, such as the extent to which a word is discussed in contexts about women versus men. This approach is generalizable to a range of key terms and dimensions, making it useful to many sociological domains.

As one example of a sociological application, Jones et al. used this approach to investigate the gendered associations of words about career, family, science, and art, using word embeddings trained across two centuries of books (2020). Their findings suggest that many of the gender stereotypes in these domains have receded in language over time. As another example, Kozlowski et al. used word embeddings to investigate five dimensions of meanings relevant to social class in books across the twentieth century (2019). Their results suggested that the cultural associations of education with social class emerged only in recent decades, while in the earlier part of the 20th century these associations were mediated by meanings of cultivation. As these studies illustrate, word embeddings enable social scientists to investigate cultural phenomena in ways that may be impractical or even impossible using surveys or other traditional social science methods.

More broadly, social scientists often use word embedding methods to investigate the relationships between language, widely held personal meanings (e.g., from survey responses), and broader societal patterns (e.g., demographic trends). For instance, Caliskan et al. illustrated the correspondences between various implicit associations in human participants[2] and cosine similarities between corresponding word vectors (2017; see also Lewis and Lupyan 2020). Further, Garg et al. showed that the way occupations are gendered in word embeddings corresponds tightly (correlations around 0.9) to the proportion of women in these occupations based on Census data, both in the present day and across time (2018). Both papers identified these patterns across embeddings trained on a variety of media corpora. The results both validated word embedding methods and illustrated the surprising extent to which our language encodes undesirable biases and inequalities. Given that embedding methods enable social scientists to efficiently leverage historical text data, a stream of social science uses word embeddings to specifically investigate cultural change (e.g., Best and Arseniev-Koehler 2022; Boutyline et al. 2020; Jones et al. 2020; Kozlowski et al. 2019; Rozado and al-Gharbi 2021). Thus far, this work has been largely associative, but it paves the way for future, more causal work on the links between language, culture, and material patterns.

Despite the promise and prevalence of word embeddings for *empirical* research, sociologists are just beginning to reconcile these methods with sociological notions of meaning and culture (e.g., Arseniev-Koehler and Foster 2020; Kozlowski et al. 2019). To theorize word embedding methods, it is important to first understand how we arrive at a trained word embedding from raw text data (i.e., the training data). These methodological details are briefly reviewed next.

[2] Measured with the implicit associations test (Greenwald, McGhee, and Schwartz 1998).

Approaches to Compute Word Embeddings Commonly used in Social Science

Social scientists use a wide variety of algorithms to estimate (i.e., train) word embeddings from text data. In this paper, I focus on describing one key algorithmic difference which will be most relevant to my theoretical arguments: the use of count-based approaches or using a machine-learning framework called artificial neural networks (i.e., neural word embeddings) to train word embeddings from text data (Baroni et al. 2014).

Count-based approaches begin with a word by word (or word by document) co-occurrence matrix computed from the entire corpus and attempt to reduce the dimensionality of this matrix by finding *N* latent, lower-dimensional features that encode most of the matrix's structure. A wide variety of methods have been used for dimensionality reduction. The output from performing dimensionality reduction is a vocabulary size by *N* dimensional matrix: each row is an *N*-dimensional word vector. Among the most popular and successful count-based word embedding approaches is GloVe (Pennington et al. 2014). However, since the publication of word2vec (Mikolov, Chen, et al. 2013), neural word embedding architectures are becoming dominant in computer science for their flexibility and performance on downstream tasks (Baroni et al. 2014; Mandera et al. 2017).[3]

[3] Word2vec remains among the most popular and parsimonious neural word embedding models used in social science, and thus is one core neural word embedding model referenced in this paper. However, computer scientists have also developed a wide variety of neural word embeddings which have specialized features. Some of these variants are described as relevant in later sections (e.g., "contextualized" embeddings and multimodal word embeddings).

Neural word embeddings use artificial neural networks to *incrementally learn* word-vectors from a given corpus as they "read" the text data and attempt to predict missing words in the data. For example, in word2vec[4] with Continuous Bag-of-Words (word2vec-CBOW), the model learns word vectors while attempting to "fill-in" missing words from various sets of contexts in the text (Mikolov, Chen, et al. 2013). Word2vec-CBOW is iteratively given a context of words (e.g., 10 words) with one word missing, and is tasked with predicting the missing word. To make a prediction, the model first combines the observed context word vectors to form a single vector representing the context. Second, it predicts the missing word based on the word vector which is *most similar* (or closest in space) to this context vector.[5] Since word vectors may be initially randomized, the model initially tends to predict the missing words incorrectly. Each time it guesses incorrectly, the correct word is revealed, and the model updates the word vectors to reduce this prediction error (and thus, improve its chances at guessing correctly if it were to see this context again).[6] As word2vec adjusts word vectors across many attempts to predict

[4] This explanation of word2vec is simplified for brevity. For example, in word2vec-CBOW, the context-vector is not merely the average of the context word-vectors, since high frequency words are downweighed in certain word2vec architectures (Arora, Li, Yingyu, et al. 2016). For details on the word2vec architecture, see Rong (2014).

[5] More precisely there are *two* word vectors for each word, one corresponding to contexts and one to target words. For details on the word2vec architecture, see Rong (2014)

[6] More specifically, the objective is to maximize the similarity between the missing word and the context words and minimize the similarity between the observed missing word and vocabulary words not in the context. Minimizing the similarity between the missing word and all vocabulary

words from their contexts, the word vectors begin to better represent words. Upon reaching some pre-determined stopping point (e.g., the level of accuracy), social scientists often stop training and use the most recent word vectors for downstream analyses. Other than CBOW, the second possible word2vec training task is Skip-Gram, where the goal is to guess the context words around the target word.

All these variants of neural network based and count based embeddings produce a vocabulary size by *N* dimensional matrix, where each row is an *N*-dimensional word vector, and perform similarly on linguistic tasks (e.g., Mandera et al. 2017). Still they offer very distinct mechanistic explanations for learning and processing linguistic information (e.g., Arseniev-Koehler and Foster 2020; Günther et al. 2019; Hollis 2017; Mandera et al. 2017).[7] As I will

words not in the context is impractical to implement. Therefore, in practice this is often approximated by minimizing the similarity between the missing word and *k* other words not in the vocabulary (i.e., negative sampling).

[7] Count-based and neural word embeddings have been shown to perform comparably on semantic tasks such as completing analogies and evaluating word similarity (Levy and Goldberg 2014; Levy, Goldberg, and Dagan 2015). In theory, these embedding approaches are also trying to extract very similar information from the raw text data. Both GloVE and word2vec (with skip-gram, negative sampling, and certain parameter settings) have been shown to being doing implicit matrix factorization (Levy and Goldberg 2014). Still, the algorithmic difference between them is crucial to their interpretation and theorization (Baroni, Dinu, and Kruszewsk 2014; Günther, Rinaldi, and Marelli 2019; Mandera, Keuleers, and Brysbaert 2017). For instance, predictive and count-based approaches provide very different mechanistic explanations for how

show, these differences have implications for theorizing the kind of meaning that word embeddings operationalize.

Computer scientists developed contemporary word embeddings to enable computers to learn, process, and represent human language, not to operationalize any particular theory of language or linguistic meaning.[8] Now, these methods are rapidly gaining traction in social science and serve as a foundation for many language modeling advances in computer science. Therefore, it is crucial that we clarify what they do and do not operationalize. This paper critically evaluates the possibility that word embeddings operationalize structural linguistics.

A PRIMER ON STRUCTURAL LINGUISTICS

Structural linguistics is both a theoretical perspective on language, and an analytic approach to study language (Craib 1992:131–48; Joas and Knobl 2009:339–70; Leschziner and Brett 2021). In this paper, I focus on three premises of structural linguistics which were influential in cultural

meaning may be learned and processed, and thus operationalize slightly different notions of meaning (see also Arseniev-Koehler and Foster 2020; Günther et al. 2019). Analogously, two different agent-based models may arrive at similar macro-level outcomes, even if they have different assumptions about what agents do.

[8] While contemporary word embeddings were developed in computer science for the purpose of quantifying language, rather than operationalizing any theory of meaning, they imported approaches used in cognitive science to quantitatively model how humans process and represent language (e.g., Latent Semantic Analysis, (Landauer and Dumais 1997)).

sociology. A first core premise is that language is relational: it is comprised of various signs (e.g., words, suffixes, phonemes) and these signs are defined by their *relationship* to other signs, rather than by any external reality (Saussure, 1983, p. 113). For example, a word is defined by its co-occurrence relationship to all other words – *not* from the intrinsic properties of the letters or sounds that comprise the word, from dictionary definitions, or by its reference to some external object. This suggests, for example, that if a misspelled word is *used* in a similar way as a correctly spelled word, both spelling variants will be understood in the same way. If, however, spelling variants are used in some systematically different way (e.g., British versus American spellings), the variants may be understood as distinct– even when both spellings refer to the same physical object. Envisioning language as relational, rather than rooted in referents, may also be interpreted as seeing language (or any other symbolic system) as self-contained and autonomous (Barthes 1977).

Identifying and understanding the relational structures in language is the core goal of structural linguistics. One well-studied type of relationship is a binary opposition (Lévi-Strauss, 1963, p. 35): a structure of meaning where two concepts are defined by their oppositional relationship to one another. In this perspective, for example, we cannot conceive of the concept of "good" without that of "evil" because they form a binary opposition. "Good" is defined by its distinction from "evil" and vice versa. Theoretically, structuralism suggests that binary opposition is a key, latent structure of meaning which scaffold symbolic systems, such as language. Therefore, a common empirical goal in structural linguistics (and structuralist-inspired scholarship more broadly) is to identify binary opposition (e.g., Alexander and Smith 1993; Barthes 2012; Jones and Smith 2001; Lembo and Martin 2021; Lévi-Strauss 1963:35).

A second core premise of structural linguistics is that underlying the inconsistent ways in which we use words ("parole"), there exists a latent, stable, and *coherent* linguistic system ("langue"). Structural linguistics focuses on studying langue rather than the varying ways in which we use words (or other linguistic units). Similarly, contemporary work influenced by structural linguistics also often hypothesizes and studies a latent system organizing the disparate ways in which a set of symbols are used (e.g., Barthes 1961; Cerulo 1995; Tavory and Swidler 2009).

Initially, langue was also described as something psychological (i.e., internalized), and generalized beyond any individual language user (thus shared or cultural) (Saussure 1983; see also Stoltz 2019). However, scholarship influenced by structural linguistics remains divided between envisioning symbolic systems external to individuals' minds (i.e., in public culture) versus as something internalized or cognitive (e.g., Alexander 2003; Lévi-Strauss 1963). As I will illustrate at various points in this paper, word embedding methods have also inherited this division; alternately being used to study meaning in personal vs public culture (Lizardo 2017).

A third core premise of structural linguistics is the distinction between studying language as a static versus dynamic system. In structuralist jargon, these are referred to as "synchronic" versus "diachronic" analyses. The former considers how the parts within a linguistic system interact at any given time point (e.g., what are the kinds of the relationships that exist between words or morphemes). The latter focuses on how this system changes and why (e.g., how words' positions in the system change across time or how new words emerge). Analogously, one can study chess as a static or dynamic system: we can "freeze" a chess game and describe where the pieces lie on the chess board in relation to one another, or we can describe the movements of pieces across a game (Saussure 1983). Structural linguistics focuses on theoretically understanding and empirically studying language as if static.

OPERATIONALIZING STRUCTURAL LINGUISTICS WITH WORD EMBEDDINGS

Here, I detail the ways in which word embeddings can be used to operationalize each of the three premises of structural linguistics described previously: the focus on language as a relational, coherent, and static system. Where relevant, I distinguish between cognitive and non-cognitive interpretations of these premises which, as noted earlier, are two variants of structuralism. I primarily consider the three premises as they pertain to the meanings of words. However, given that structural linguistics generalizes to linguistic units other than words (e.g., morphemes) and to symbolic systems beyond language, many of the following arguments may be widely generalized as well.

Modeling Language as a Relational System with Word Embeddings

Word embeddings operationalize language as a relational system in several ways. Most crucially, like many text analysis methods, they rely on the Distributional hypothesis (Firth 1957; Harris 1954). This hypothesis suggests that words may be understood by differences in “the company they keep” – i.e., their co-occurrence relations. It is no accident that this hypothesis is fundamentally relational: It emerged in structural linguistics (Sahlgren 2008), not computer science.

While all word embeddings operationalize the Distributional hypothesis, they do so in radically different ways. Count-based models learn from global patterns of co-occurrence, while neural network models learn from many local contexts. The fact that both approaches work comparably is remarkable, illustrating how global relationships between words can be estimated from many local contexts. Analysts also vary widely in how they define these contexts (e.g., the size of the context window, and how they combine context words into a single context vector). In all cases, the resulting word-vectors are also only defined relationally: They are only interpretable to human analysts because of their *relative* positions. Word vectors are not tied to any external referents, and any given word vector is arbitrary and uninformative outside of its semantic space.

The extent to which word embeddings operationalize a relational theory of *meaning* depends on the analyst's interpretation of "meaning" in the Distributional hypothesis. Indeed, the hypothesis is notoriously vague when it comes to the relationship between distributional structure and meaning (Harris 1954:151–57; see also Sahlgren 2008). The concept of meaning is only relevant when the analyst introduces it. In practice, researchers often implicitly interpret meaning in the Distributional hypothesis somewhere along two extremes (Lenci 2008).

At one extreme lies the weakest reading of the hypothesis: A word's meaning – whatever that might be – *correlates* to the patterned ways in which it is used in language (e.g., Harris 1954). In this first reading, "meaning" is latent: Word embeddings are not necessarily capturing any meaning at all, let alone a relational notion of meaning. Under this interpretation, word embeddings trained on large-scale cultural texts offer *proxies* for meaning. This use of word embeddings may also be motivated by the intuition that public culture, like media and books, *reflects* personal culture (e.g., Garg et al. 2018; Kozlowski et al. 2019; Xu et al. 2019).

At the other extreme, in a stronger interpretation of the hypothesis, the meaning of a word *is* based on its distributional patterns in natural language (rather than, say, its dictionary definition, the emotions evoked by a word, or a word's reference to a physical object). This second interpretation is distinctly structuralist, suggesting that words are defined relationally rather than by anything external to the linguistic system. This second interpretation may also be cognitive and causal, suggesting that our meanings of words may be learned from (and influenced by) the relational patterns in natural language. A midrange ("partly structuralist") interpretation is that meaning is, in part, learned from the relational pattern of words. Under this cognitive interpretation of the distributional hypothesis, word embeddings measure (at least to some extent) the "meanings" that may be evoked by language. Word embeddings, then, are empirically useful to investigate the meanings that can be learned from, or reinforced by specific language sources, like children's books (Lewis et al. 2022) or news reporting on obesity (Arseniev-Koehler and Foster 2020). More generally, this approach to using word embeddings is often motivated by the intuition that public culture *shapes* personal culture.

Finally, empirical analyses using word embeddings often focus on identifying relational structures – primarily, binary oppositions (e.g., Arseniev-Koehler and Foster 2020; Best and Arseniev-Koehler 2022; Boutyline et al. 2020; Caliskan et al. 2017; Garg et al. 2018; Jones et al. 2020; Kozlowski et al. 2019; Nelson 2021; Taylor and Stoltz 2021). For instance, as described in part one, a core approach to measure a concept, like gender, is to identify a line between the word vectors for two opposing poles (e.g., "woman" and "man" for gender). This measure operationalizes the concept of gender as ranging continuously from one pole to the other. Being closer to one pole implies being farther from the other (e.g., more masculinity implies less femininity). In a similar approach, analysts identify word vectors corresponding to two poles of an opposition (e.g., “woman” and “man” to represent gender) and then compare the distance of some interesting word to each pole. This second strategy does not assume that more femininity implies less masculinity or that gender is represented as a line in semantic space, but still identifies concepts as oriented by two poles and thus assumes they are relational constructs.

Modeling Language as a Coherent System with Word Embeddings

From the varying ways in which words are used in a corpus, word embeddings aim to abstract a latent, coherent system. This may be thought of as abstracting “langue” from “parole.” The abstraction occurs in several ways. First, word embedding methods commonly used in social science (e.g., word2vec) represent each word as a single word vector. Thus, the goal is to capture the regularities of words across all the various contexts in which the word appears in the corpus. Modeling each vocabulary word as a vector assumes there is some degree of systematicity across the varied instances of each vocabulary word in the data.

Second, the architecture of word embeddings aims to find latent regularities *across* vocabulary words, such that a limited number of dimensions can represent all words in a vocabulary. More specifically, words are represented as vectors where each element corresponds to a loading on each of *N* dimensions, as described in the section earlier *A Primer on Word Embeddings*. The dimensionality (*N)* of word vectors is always far lower than the vocabulary size of the corpus. Often, *N* is set at a few hundred, while vocabulary sizes often range from tens of thousands to several hundred thousand words depending on the corpus. This difference in sizes assumes that there are latent patterns across the raw co-occurrences of words which will accurately capture a high dimensional vocabulary. Dimensions are shared and reused across vocabulary words to represent different aspects of their meaning, and so only a limited number of dimensions is needed to model words. In fact, this compression is thought to be critical to the performance of word embeddings compared to word-vectors based on raw co-occurrences (Arora, Li, Yingyu, et al. 2016). The high performance of trained word embeddings (e.g., at capturing human-like semantic information) also attests to the validity of modeling words' meanings as coherent.

As noted in part two, interpretations of langue diverge as to whether it is internal or external to individuals. Similarly, current social science employing word embeddings also comes in a cognitive and non-cognitive flavor. Some work uses word embeddings as cognitive models: to learn about how semantic information might be learned by, represented in, and processed by human minds (e.g., Arseniev-Koehler and Foster 2020; Baroni et al. 2014; Günther et al. 2019). Other work uses embeddings as methods to learn about cultural texts and broad-scale cultural change (e.g., Garg et al. 2018; Kozlowski et al. 2019). These two approaches to using word embeddings may be thought of as operationalizing cognitive and non-cognitive interpretations of langue. The former aims to abstract langue as a system that is internalized and *learned from* cultural texts, while the latter aims to abstract langue as a latent system of meaning *encoded in* a cultural text.

Modeling Language as a Static System with Word Embeddings

Word embeddings can also model language as a static system of signs, following a structural linguistic perspective. Neural and count-based word embeddings do so differently. As described earlier, count-based embeddings abstract a semantic space from global co-occurrences while neural word embeddings learn a semantic space incrementally, as the by-product of "reading" and predicting data. When social scientists *use* word vectors from neural embeddings for empirical analyses, they pause the training[9] and extract out the word-vectors as a semantic space. Thus, all word embedding can be used to examine language as a static system, but count-based and neural word embeddings do so in different ways. As I will elaborate later, this difference matters for its implications on how word embeddings hold up to critiques of this structuralist premise.

Social scientific analyses using word embeddings vary in the extent to which they focus on language as static or dynamic. While some empirical work using embeddings investigates a cultural phenomenon at a single time point (e.g., Caliskan et al. 2017; Lewis and Lupyan 2020; Nelson 2021), a large body of scholarship uses word embeddings to track sets of words or concepts (e.g., stereotypes) across time (e.g., Best and Arseniev-Koehler 2022; Boutyline et al. 2020; Hamilton et al. 2016; Jones et al. 2020; Kozlowski et al. 2019). The former's static lens may be considered more characteristically "structuralist."

FOUR CRITIQUES OF STRUCTURAL LINGUISTICS

Here, I consider four critiques of structural linguistics: (1) that meaning may be grounded or embodied, rather than purely relational, (2) that a focus on binary oppositions is reductionistic,

[9] In practice, we usually stop the training process based on preset hyperparameters in the algorithm – such as error falling below a certain threshold or following a set cap on the number of iterations of the algorithm.

(3) that meaning is incoherent, and (4) that language is dynamic. Like the previous section, I focus on these critiques as they relate to language specifically, recognizing that many of these critiques and the following arguments generalize beyond words and language to other symbols and semiotic systems. After briefly introducing each critique, I argue how it applies (or does not apply) to word embeddings and highlight implications for sociological applications of word embedding methods.

Critiques of Purely Relational Approaches to Meaning

While structural linguistics theorizes signs as defined by their relationships with one another, scholarship also highlights that our understandings of words are linked to concrete referents in the external world: physical objects and events, sensorimotor experiences, and/or emotional experiences (e.g., Barsalou 1999; Lakoff and Johnson 2008; Moseley et al. 2012; Pulvermüller 2013; Quiroga et al. 2005; Smith and Gasser 2005). For example, like word embeddings, humans can know when and how to use the word "summer" next to words like "spring" and "sun." But we can also understand when "summer" refers to a specific, upcoming time point. And we can identify other, non-linguistic references to "summer," such as from a calendar. These examples illustrate that meaning may be *grounded* in concrete referents (see also Bryson 2008). We may also learn the meanings of "summer" as something about relaxation, joy, sunshine, and warmth, from our experiences in summer months. Hearing or seeing the word may then evoke these same sensations and feelings; perhaps our meaning of the word "summer" entirely consists of the feelings and sensations evoked by the word. These examples illustrate that meaning may be *embodied* – that is, linked to our bodily sensations.

Across the disciplines, there is growing consensus that our brains incorporate semantic information from a variety of sources (Davis and Yee 2021; Quiroga et al. 2005), not merely from language. However, exactly how this occurs is less understood. For instance, it is unclear to what extent meaning is linguistic versus embodied/grounded, and how semantic information from distinct sources (e.g.., linguistic and sensorimotor) is combined. Therefore, this critique highlights that while the meanings of words (and other symbols) are relational to some extent, language is far more than just a self-contained system of signs.

Implications of these critiques for word embeddings.

The extent to which critiques of a relational approach to meaning apply to word embeddings partly depends on the analysts' interpretation of "meaning" in word embeddings. As detailed in the section on *Modeling Language as a Relational System*, a weaker reading of the Distributional hypothesis is that a word's meaning correlates to its relationship to other words in a language. A stronger reading of the Distributional hypothesis is that a word's meaning *is* defined by its relationship to other words in a language. This stronger interpretation is also more structuralist, and vulnerable to broader critiques that linguistic meaning may also be grounded or embodied, rather than merely relational.

Despite critiques of a purely relational notion of meaning, some words do not have concrete references or are not easily experienced and thus are unlikely to be learned from concrete referents or sensorimotor information alone (Borghi et al. 2017). For example, not all humans have experienced the words "dive," "depression," or "royalty," and yet we know exactly how to use these words to transmit and build up larger ideas. More generally, it is unlikely that we learn the meanings of abstract words like “epistemic” and “subjective” from sensorimotor experience or physical objects. This suggests that meaning may not be entirely embodied, either. Indeed, the Distributional hypothesis suggests a *mechanism* for learning and communicating more abstract concepts: the relational patterns of words in language (Borghi et al. 2017; Günther et al. 2019).

In fact, the high performance of word embeddings across a range of linguistic tasks (e.g., solving analogies) provides one of the most convincing *demonstrations* of the Distributional hypothesis, and a relational notion of meaning more generally. Even though word embeddings learn from the distributional patterns of words alone, they learn semantic information that strongly correlates to what humans learn (for a review, see Caliskan and Lewis 2022). This suggests that a mid-range reading of the Distributional hypothesis is warranted: that meaning correlates to distributional patterns of words and meaning may be also partly learned from distributional patterns in language (see also Davis and Yee 2021; and Lenci 2018).

Even if word embeddings operationalize meaning as purely relational, they can still be very useful for sociologists to study sensorimotor and emotional information that is encoded into language. For instance, research might test how language about senses is used to make sense of other domains, like descriptions of sexual relationships in terms of sweetness, bitterness, heat, or cold (see also Tavory and Swidler 2009). Dictionaries with sensorimotor information about words may be used to identify words about senses (Lynott et al. 2020). More generally, researchers can test for a range of conceptual metaphors in language (Lakoff and Johnson 1980), such as how semantic information about orientation organizes semantic information about morality (Lakoff and Johnson 2008). Because word embeddings encode rich semantic representations (and are scalable), they can be used to address calls to consider embodied knowledge as part of cultural and cognitive sociology (Ignatow 2007), especially through text analysis (Cerulo 2019; Ignatow 2009, 2016).

Further, a stream of work in computer science aims to develop language models where meaning is *both* relational (learned from distributional patterns in text data), and is learned from extra-linguistic experiences, such as images of what a word represents (Baroni 2016; Bruni, Tran, and Baroni 2014; Goh et al. 2021; Li and Gauthier 2017; Radford et al. 2021; Roy 2005; Vijayakumar, Vedantam, and Parikh 2017). These multimodal word embeddings integrate semantic information derived from text, images, sound, or other modalities. Empirically, multimodal embeddings capture slightly different information than may be learned from text alone (e.g., Vijayakumar et al. 2017). For example, while word2vec learns the word "apple" as closest in space to "apples," "pear," "fruit," and "berry," a word2vec model also trained on sounds learns "apple" as closest to "bite," "snack," "chips," and "chew" (Vijayakumar et al. 2017). Multimodal embeddings are not yet popular in social science but offer an exciting direction for sociologists to address and overcome critiques of a relational notion of meaning.

Critiques of Binary Oppositions

The focus on binary oppositions in structural linguistics (and structuralism more broadly) also garners extensive critique (Craib 1992). Binary oppositions are one among many possible forms of meaning. Opposition itself comes in many varieties: hierarchical, continuous, dichotomous, or graded (Geeraerts 2010:87). For instance, we might describe aesthetics as a dichotomous concept (as unattractive versus beautiful), or on a graded scale (unattractive versus plain versus pleasant versus beautiful). Oppositions might be discrete and mutually exclusive, such that one pole necessitates the complete lack of the other (e.g., dead versus alive). They may have an evaluative component, such as good versus bad and clean versus dirty. We might also have ensembles of multiple oppositions. For example, the Western meaning-system for direction consists of two binary oppositions (north/south and east/west) or of three oppositions (up/down, left/right, and forward/backward) (Geeraerts 2010:87). Finally, concepts may be multidimensional, such as the constructions of race and ethnicity. In sum, the structuralist focus on *binary oppositions* (as opposed to other oppositions or other forms of meaning) may be overly reductionistic.

Implications of these critiques for word embeddings.

Limitations of binary oppositions in structural linguistics also directly apply to a body of work in word embeddings which also focuses on binary oppositions. Indeed, analysts have measured a wide variety of concepts as binary oppositions in semantic space, such as gender, age, and size. At the same time, scholars find tremendous variation in the extent to which the resulting measures actually match human-rated perceptions of the concept (e.g., Chen, Peterson, and Griffiths 2017; Grand et al. 2018; Joseph and Morgan 2020). These findings underscore theoretical critiques of the limitations of binary oppositions.

Gender is the canonical case for studying concepts as binary oppositions in word embeddings. But gendered meanings in word embeddings (when measured as a binary opposition) also appear to have an especially tight correspondence to human-ratings – unlike, for example, race (Grand et al. 2018; Joseph and Morgan 2020). Perhaps, gender is also an outlier in the extent to which it manifests in raw text data as an opposition between two poles (see also Ethayarajh, Duvenaud, and Hirst 2019). Indeed, gender is frequently and explicitly denoted in language (as masculine vs feminine) with pronouns, suffixes, and other grammatical endings. Alternatively, perhaps human raters organized the concept of gender around two poles, more so than they tended to do for other concepts. Indeed, gender is pervasively constructed as a binary between men and women (Ridgeway 2011). Although this work cannot yet resolve why gender is an outlier, it does highlight limitations and nuances of measuring just any concept as a binary opposition using word embeddings.

Some scholarship goes beyond binary oppositions – looking for *systems* of oppositions and other latent structures in space. Kozlowski et al. investigated class as a system of oppositions – investigating the relationships between five dimensions of class across time (2019). Boutyline et al. investigated gendered stereotypes relevant to education in print media from 1930-2009, including the gendered cultural associations of effort and intelligence (stereotypically feminine and masculine routes to success, respectively). Across time, the gendered associations of effort and intelligence became increasingly and synchronously polarized: as the former gained feminine associations the latter gained masculine associations. These results suggest that these gendered stereotypes changed together across time as an opposition to one another (2020).

Finally, researchers have also begun to investigate information structures *beyond* oppositions, such as topical regions or clusters of words in semantic space (e.g., Arora et al. 2018; Arseniev-Koehler et al. 2022). As demonstrated by Nelson (2021), because word-vectors can be decomposed and recombined, word embeddings can also be used to look at meaning from an intersectional lens where binary oppositions may *interact*. This can be done, for instance, by combining the word vectors “woman” and “Black,” and comparing this with the combination of “woman” and “white.” Meanwhile, another body of work investigates how to build word embeddings that can model even more nuanced forms of information, such as hierarchy (Nickel and Kiela 2017). Thus, while early scholarship using word embeddings heavily focused on binary oppositions, a stream of emerging scholarship considers other structures that can overcome critiques of binary oppositions.

Critiques of Coherence

The coherence posited by structural linguistics (and structuralism more generally) is also one of its most controversial aspects. From this perspective, structuralism envisions meaning (of words or other symbols) as unrealistically logical and homogenous (e.g., Bakhtin 1981; DiMaggio 1997; Martin 2010; Sewell 2005:169–72; Swidler 1986, 2013). Scholars emphasized the necessity of accounting for context (e.g., Douglas 2003; Geertz 1973; Labov 1972). In the case of language, a word’s meaning may vary depending on a host of factors, such as where the text is produced and by whom, or who is reading the text (e.g., Franco et al. 2019; Geeraerts, Grondelaers, and Bakema 2012; Geeraerts and Speelman 2010; Hu et al. 2017; Peirsman, Geeraerts, and Speelman 2010; Robinson 2010).

As an example, the word “awesome” evokes multiple distinct interpretations in different linguistic contexts; certain individuals are more likely to interpret “awesome” according to some of these interpretations than others, depending on their age, gender, and even neighborhood (Robinson 2010). Even among uses of a word within a given document, the word may evoke very different interpretations depending on its surrounding words — a phenomenon known as polysemy. The word “depression” has entirely different meanings in a sentence about mental health versus one about economics. The word “depression [economics]” is even more specific the context of “The Great Depression,” which refers to a particular economic depression. Such examples of the variation of words highlights the shortcomings of focusing on language as insensitive to context.

Implications of these critiques for word embeddings.

Word embedding methods commonly used in social science, such as word2vec and GloVe, are vulnerable to the same longstanding critiques of coherence as structural linguistics. Most crucially, these models use a single word vector for each vocabulary word in the corpus, thus smoothing over the variable ways in which a word is used across the training corpus. Because these models are insensitive to linguistic context, they are commonly critiqued as modeling words' meaning as unrealistically coherent (e.g., Faruqui et al. 2016; Gladkova and Drozd 2016; Neelakantan et al. 2015; Wang et al. 2020).

In fact, this limitation is well-known in computer science, and prompted a variety of approaches to allow for linguistic meaning to be messier and more context dependent. Most notably, computer scientists developed a new paradigm to model language in computer science: "contextualized" neural word embeddings (Devlin et al. 2019; M. Peters et al. 2018). While models like word2vec and GloVe represent each vocabulary word as a vector, contextualized models produce a vector for each instance of a word in a text. In a contextualized embedding, for example, each time the word "depression" is used in a corpus it may be modeled with a slightly different vector. The raw word-vector for "depression" is modified based on the context words used around each mention of "depression." Thus, contextualized models enable words to vary across linguistic contexts.

The specific approaches to contextualize word vectors vary widely, but all use some form of an artificial neural network (for reviews, see M. E. Peters et al. 2018; Wang et al. 2020). The broad goal of training is to learn both stable aspects of language and contextual aspects of language. To gain intuition into how a model might contextualize word-vectors, consider a very simplified strategy (roughly based on Akbik, Blythe, and Vollgraf 2018; and Peters et al. 2017).. Like word2vec, an artificial neural network is tasked with "reading" a sentence and predicting the next word. However, as this network predicts the next word in the sentence, it also keeps an ongoing vector representing the "gist" of what is currently being talked about at any point in the sentence.[10] This "gist" is updated with each new word encountered in the sentence, and it is a function of the sequence of preceding words. Part of the model's training process is learning how to maintain this "gist": learning what information to keep, what to forget, and how to use information previously encountered, as it reads a sentence and predicts a word. Once training finishes, we can input a sentence and, for any word used at some point *t* in the sentence, we can extract out the "gist" at time *t*. This "gist" is still a vector but it is a function of the preceding context words in the sentence. This gist may also be combined with a non-contextualized word vector (e.g., concatenated or summed). Regardless of the approach, contextualized word embeddings allow for heterogeneity across linguistic contexts.

Still, even contextualized word embeddings do not entirely overcome critiques of coherence. For instance, they still do not account for heterogeneity across extra-linguistic contexts — such as who produced the text, when, where, or why. In addition, contextualized

[10] In more technical jargon, this conceptual description refers to the hidden state in a long-short-term memory (LSTM) network.

models (particularly the most recent ones, like BERT (Devlin et al. 2019)) require extraordinarily large corpora for training. Therefore, training from scratch on a dataset is generally impractical or impossible. Instead, researchers typically begin with one of a few, select models (sometimes called “foundation models”) which are already trained on supersized corpora (Bommasani et al. 2021). They then continue to train (i.e., “fine tune”) the model to their specific data or task. While foundation models are remarkably adaptable, even after fine-tuning they will still reflect their initial training data in various ways (Merchant et al. 2020). Thus, even these models assume there is some underlying coherence to linguistic meaning.

Using traditional (i.e., non-contextualized) word embeddings, one practical strategy to address the inconsistent uses of words within a corpus is to train multiple models on various subsamples of the data (e.g., bootstrapping). This strategy is often used to ensure that findings are robust to any particular subset of documents (or other contexts) (e.g., Boutyline et al. 2020; Kozlowski et al. 2019). But it also reveals how sensitive empirical findings are to specific usages of words. Resampled or bootstrapped embeddings make it possible to model variation of a words’ meanings across the contexts. They also make it easier to distinguish between patterns that are robust across documents versus specific to subsets of documents. However, this approach is not as sensitive to context as contextualized word embeddings, and it still does not account for the many other sources for variation in linguistic meaning.

For sociological applications of word embeddings, contextualized models and resampling/bootstrapping offer ways to move past a dichotomous view of cultural meaning as either coherent or incoherent (Ghaziani 2011). Indeed, a recent study in computer science investigated the extent to which meaning is contextual in contextualized embeddings, by comparing words’ vectors from contextualized and non-contextualized embeddings (Ethayarajh

2019). This study found that, on average, less than 5% of the meaning of a word's contextualized word vectors (where there is one word vector from each instance of the word in the corpus) could be explained by a single word vector. Further, contextualized models generally outperformed non-contextualized models on various linguistic tasks. At first glance, these results might suggest that enabling meaning to be incoherent more closely models human meaning, or that a coherent model of language is indeed overly unrealistic. However, the extent to which contextualized embeddings outperform static models in downstream applications varies widely across specific linguistic tasks (Arora et al. 2020; Ethayarajh 2019; Tenney et al. 2019). Indeed, for many tasks and corpora, contextualized word embeddings only yields only marginal improvements (Arora et al. 2020; Tenney et al. 2019). In some of these cases, contextualized and non-contextualized embeddings even perform equally, thus *validating* structuralist visions of coherence.[11] For social scientists, these findings suggest that the extent to which meaning is coherent is far more nuanced and remains an open (and promising) research area well suited to word embedding methods.

[11] Further, empirical work illustrates it is not necessarily the *contextualization* process that leads to contextualized models' improvement on linguistic tasks (Arora et al. 2020) Indeed, contextualized embeddings also often include many other training architectures such as predicting sentences as well as words, accounting for suffixes and affixes, and accounting for the order of words. Compared to models like GloVe and word2vec, contextualized models also have enormous numbers of hyperparameters (i.e., knobs to tune to transform the inputted context words) and are trained on much larger corpora.

Critiques of a Static Lens on Language

Structural linguistics conceptually distinguished the study of language across time from the study of language at a single time point. While useful for analytical purposes, it also struggled to ever reconcile these two lenses (Giddens 1979:13; Stoltz 2019). Even if we give precedence to theorizing a symbolic system at a single time point, we also need to be able to explain changes in this system (Emirbayer 2004:10–11; Giddens 1979). As a theory, and even as a framework for empirical analysis, structural linguistics cannot account for how language (or any other symbolic system) may be both static and dynamic.

Implications of these critiques for word embeddings.

The extent to which word embeddings are vulnerable to critiques of a static lens partly depends on the approach used to learn word embeddings: count-based versus artificial neural network based. Count-based word embeddings model a symbolic system as static but abstract the whole semantic space at once by performing some dimensionality reduction on a co-occurrence matrix, as described in section one. These methods do not incorporate any mechanism for change in a semantic space. Thus, count-based embeddings, like structuralist linguistics, cannot reconcile a static and dynamic account of language.

By contrast, *neural* word embeddings (including contextualized models) model language more dynamically. The word vectors are deployed and updated each time new cultural stimuli (e.g., text excerpts) are encountered during training. Upon experiencing a context (i.e., text excerpt), the neural word embedding uses its current information about each vocabulary word (i.e., “looks up” the word’s position in the semantic space at this point) to make a prediction about the missing word in the context. When the prediction is incorrect, the positions of word-vectors are shifted, yielding an *updated* symbolic system. Thus, word vectors structure how neural word embeddings experience any incoming language and are simultaneously structured by new experiences with language. In this way, neural word embeddings operationalize the notion that a symbolic system is both a “thing” and a “process,” i.e., a “structuring structure” (Bourdieu 1984; Giddens 1979; Sewell 1992). The symbolic system captured by neural word embeddings is part of a dynamic process, and changes as the embedding interacts with its cultural environment and experiences new data.

When we use word vectors from neural word embeddings in social science applications, we generally stop the training process and begin analyses on the “frozen” system. We have extracted the word vectors as static representations from a system that can hypothetically change at any time with additional stimuli (i.e., additional text data) if we were to “unfreeze” the system. Thus, unlike count-based embeddings, neural word embeddings lend themselves to static analyses, but do not entirely divorce static and dynamic lenses.

Importantly, this account of neural word embeddings as dynamic makes more sense for the cognitive flavor of structuralism, where langue is internalized in a single individual. “Training,” then, may be thought of as cognitive socialization (Arseniev-Koehler and Foster 2020). Notably, neural word embedding models only one possible source for meaning change within an individual: Mew experiences with cultural symbols. It does not account for many other possible factors such as social relationships, and it does not offer an account for macro-level change in meaning.

Further, while neural word embeddings offer a possible theoretical reconciliation between static and dynamic lenses, methods to *empirically* study semantic change with embeddings remain limited (for a review, see Kutuzov et al. 2018). This limitation is important to address given the growing sociological interest in investigating culture across larger time-scales using word embeddings (e.g., Best and Arseniev-Koehler 2022; Boutyline et al. 2020; Jones et al. 2020; Kozlowski et al. 2019; Stoltz and Taylor 2020). One popular approach to study semantic change is to divide up the corpus into time segments and then compare embeddings trained on these separate segments (using count based or neural word embeddings) (Kulkarni et al. 2015). Then, to compare word vectors across time points, researchers may either (1) rotate embeddings from different segments so that their word vectors are directly comparable, or (2) compare cosine similarities (i.e., between words or sets of words) in different segments. A downside of the second approach is that it assumes *most* words did not shift in meaning, and so local relationships are static. A downside of both approaches is that they require training an embedding on each time segment. As a result, they may be unfeasible for many corpora sizes of sociological interest (but see Boutyline et al. 2020). Even on large corpora this approach does not allow for very granular time segments.

A second approach is to estimate a word-vector from each time period in the corpus at once in a single, modified neural word embedding model (e.g., Bamler and Mandt 2017; Rosenfeld and Erk 2018; Yao et al. 2018). For instance, in addition to aiming to maximize the similarity of words that occur (and minimizing the similarities of words that do not), Yao et al (2018) suggest a modified model that aims to maximize the similarity of vectors for the same word which occurs at different time points. Rosenfeld and Erk (2018) propose to (1) learn time-invariant embeddings for each word, and (2) embeddings for each time point, and then (3) *combine* these using a learned function (e.g., a weighted combination) to arrive at a time-stamped word-vector. Although not motivated by any theoretical model for semantic change and more complex than prior approaches, these modified models offer exciting opportunities to investigate meaning change even in smaller-scale data and/or at more granularity.

A third approach to study semantic change using embeddings is to train neural word embeddings on documents ordered across time: “freezing” and saving the system at various points in time, and then comparing the frozen models across this training (Kim et al. 2014). This approach more clearly operationalizes how a symbolic system may change within an individual as they experience text, rather than macro-scale change. It also has practical limitations – for example, the quality of word vectors may improve with increased training, making it challenging to disentangle training effects from true changes across time using this approach. Further, words which are not present for several time points might simply appear to have no semantic change, and it is unclear how *new* words may be incorporated into the semantic system.

To sum up, neural word embedding methods offer a model of a dynamic meaning system that may be paused for static analyses. A variety of methods exist to empirically study language as a dynamic system but additional methodological work on modeling linguistic change using word embeddings is still needed. Such methods will enable social scientists to empirically analyze the structure and content of semantic systems across time in more precise and formalized ways and offer *new* insight reconciling static and dynamic lenses on language, and cultural symbolic systems more broadly.

DISCUSSION

Word embeddings open new doors for social scientists to investigate culture and language at scale. However, like any method, it is crucial that we clarify exactly what word embeddings operationalize. This paper has critically theorized how word embeddings operationalize key premises of an influential theory of language: structural linguistics. Not only can word embeddings be used to operationalize core premises of structural linguistics, their remarkable *successes* at capturing human-like semantic information attests to the validity of structural linguistic theory itself.

This paper also theorized the ways in which word embeddings succumb to or overcome several critiques of these structuralist premises, such as debates about the (in)coherence of meaning and relational notions of meaning. As highlighted in this paper, different word embedding algorithms do so differently. In general, while count-based word embeddings share many limitations of structural linguistics, neural word embeddings – and especially, contextualized neural word embeddings – offer solutions to these limitations. For example, neural word embeddings model a symbolic system as dynamic, which is then frozen for static, structuralist analyses. Further, while some word embeddings (e.g., word2vec) may model language as overly coherent, contextualized neural word embeddings offer solutions to account for variation across linguistic contexts. More broadly, theoretical shortcomings of structuralism parallel advances in computer science to address the limitations of word embeddings. This includes the move from count-based to neural embeddings, the more recent move from static to contextualized embeddings, and the growing interests in diachronic and multimodal embeddings in computer science (see also Bisk et al. 2020).

The extent to which word embeddings succumb to critiques of structuralism also depends on the analysts' own interpretation of "meaning" in word embeddings. This includes, for example, whether the analyst interprets word embeddings as a *theoretical model* or as a *method* to measure meaning. For instance, the Distributional hypothesis may be seen as a modeling one mechanism by which meaning is constructed from public culture. Or the hypothesis may be interpreted as merely an approach to capture a proxy for meaning, whatever meaning may be, thus side-stepping the concept of meaning altogether. Further, neural embeddings may be interpreted as formal models for how humans represent and process semantic information, rather than just methods to measure meaning (Arseniev-Koehler and Foster 2020). Like network analysis (Borgatti et al. 2009) and other structuralist tools, word embeddings may be understood as a method or as a metaphor (Craib 1992:133).

Directions for Future Social Science Research using Embeddings

This paper highlights numerous future research directions in computational social science and cultural sociology. First, word embedding methods open new angles into longstanding debates about the coherence of meaning. Static and contextualized word embeddings represent two ends of this debate; while the former represents meanings as entirely static, the latter represent meaning as highly sensitive to its surrounding words. Contextualized word embeddings make it possible to compare meanings across linguistic contexts. These methods thus offer strategies to measure the extent of the variation of meaning, identify patterns in the distribution of meanings (Sperber 1985), and perhaps ultimately explain how meaning may be both ordered and messy. At the same time, all word embeddings reflect their training corpus. Given that meaning may be incoherent across various possible training corpora, analysts must consider how the training corpus for any word embedding is produced, why, and by whom – and whose meanings the embedding represents or excludes. In this way, longstanding debates about coherence in cultural sociology may be relevant to contemporary ethical issues in machine-learning (Bommasani et al. 2021), such as that language models often reflect the language (and ideologies about language) of dominant social groups (Blodgett et al. 2020; Shah, Schwartz, and Hovy 2020).

Second, extrapolating from the meaning of words to the meaning of other signs and symbolic systems was an important legacy of structural linguistics. Similarly, while word embeddings focus on the meaning in written language, they generalize to a range of phenomena: from modeling nodes in a social network (e.g., Grover and Leskovec 2016) to segments in musical scores (e.g., Arronte Alvarez and Gómez-Martin 2019; Chuan et al. 2020). Many key theoretical points raised in this paper about word embeddings also extend to other modalities. For instance, all these variations of embeddings hinge on generalizing the Distributional hypothesis, where nodes, sounds, or images are defined by their relationship to other nodes, sounds, or images, respectively. One key distinction of non-linguistic embeddings is that cultural elements may not be as cleanly demarcated (unlike words in a text). Thus, non-linguistic applications revive longstanding methodological questions about what counts as a cultural element (Mohr 1998). This paper has focused on word embeddings given their recent rise in social science. However, moving beyond word embeddings to other modalities could also aid sociological research on how cultural signs, more generally, operate in our cultural environment (Bail 2014).

Third, this paper also highlighted a variety of scholarship using word embeddings to study the relationship between symbolic and material (e.g., demographic) patterns. For example, Garg et. al. studied the relationship between gendered associations of occupations in text and the gender ratios of these occupations (2018). This is an exciting direction and responds to longstanding calls to study cultural and social orders in conjunction. One especially relevant application is socio-semantics, which studies the relationships between social ties and semantic structures (Basov, Breiger, and Hellsten 2020). For instance, Linzhuo et. al. use embedding methods to study the relationship between centralization in social networks and semantic diversity (2020). Notably, this research direction also offers one more way to address critiques of the linguistic structuralist vision of meaning as a "closed" system.

Finally, while this paper considers several core premises and critiques of structural linguistics (and structuralism more broadly) as they apply to word embeddings, this intellectual movement is broad (Dosse 1997). Future theoretical work might consider how word embeddings and their specific architectures align with or diverge from these variations within structural linguistics, such as perspectives of de Saussure, Jakobson, and C.S. Peirce (e.g., Yakin and Totu 2014), and the many branches of structuralism more broadly. Future work might also consider numerous critiques of structuralism which are not covered in this paper, such as the role of agency and creativity. Such research might unveil other implicit theoretical assumptions – and potential innovations – in word embeddings.

Conclusions

Word embeddings are becoming pervasive social science approaches to analyze language, meaning, and culture using text data. However, these methods remain undertheorized. To ensure we use them effectively, it is crucial that we define what *kind* of meaning word embeddings operationalize and their implicit assumptions. Dissecting the way that word embeddings implicitly formalize (or might be used to formalize) sociological concepts can ultimately push us to redefine these concepts themselves (Merton 1948). Analogously, social network analysis pushed scholars to clarify concepts like “social tie,” “network,” and “community” (Borgatti et al. 2009). Now, word embeddings offer a new theoretical opportunity to formalize concepts in cultural sociology, such as schema (Arseniev-Koehler and Foster 2020), binary opposition (Kozlowski et al. 2019), symbolic system and symbol, and coherence.

Chapter 3. Are both policemen and policewomen police officers? The relationship between meaning and form in gendered language

ABSTRACT

Cultural symbols (e.g., words) are often described as mappings between a form (e.g., spelling or sound of a word) and meaning. The relationship between form and meaning is often described as *arbitrary* in culture sociology. However, empirical work in cognitive science and linguistics finds that words' forms (e.g., their spelling, or presence of suffixes) are often systematically related to their meanings. Most notably, survey and experimental work finds that peoples' gendered meanings of words depend on the words' gender suffixes (e.g., congress*man* vs congress*woman*). Drawing on cultural sociological work on symbols, I provide a theoretical account of how the relationship between gendered meaning and gender suffixes can be encoded in and learned from language. This paper empirically illustrates that the relationship between suffixed form and meaning is also encoded in public discourse (e.g., news reporting and web-crawled data). For scholarship and linguistic debates on gendered language, these findings resonate with the pervasive claim that public discourse is a likely source for learning gendered meanings of words. For cultural sociology, this paper critically revisits the notion of arbitrariness — a fundamental but perhaps outdated assumption about cultural symbols.

Cultural symbols are the coupling of meaning and form: for example, a word is the mapping between the word's meaning and its spelling or pronunciation (Lizardo 2016). A core premise in many sociological accounts of symbols is that their physical forms are arbitrarily linked to their meaning (Craib 1992:137; Joas and Knobl 2009:344). This premise suggests that we could not guess the meaning of a novel word, just from hearing or seeing a word by itself. It also suggests that there is no predictable or systematic relationship between words' spelling and sounds and their meanings.

However, empirical work in cognitive science and linguistics challenges this assumption (e.g., Blasi et al. 2016; Dingemanse et al. 2015; Köhler 1929; Perniss, Thompson, and Vigliocco 2010; Ramachandran and Hubbard 2001). Such work finds a range of systematic relationships between words' forms (spellings and sounds) and people's cognitive meanings of these words. In particular, gendered suffixes (e.g., police*man* and police*woman*) are consistently related to how people associate these words with women or men (Boroditsky et al. 2003; McConnell and Fazio 1996; Phillips and Boroditsky 2003; Sera, Berge, and del Castillo Pintado 1994). Such scholarship speculates that people learn this relationship between meaning and form *from the way that words are used* (Atagi et al. 2009). For instance, people learn that the word "policeman" has masculine connotations because it is used to describe men or in contexts about men, rather than women. In this paper, I outline and empirically test how this relationship between meaning and form might actually work: How can the relationship between form and meaning be encoded in, and learned from, language?

First, I offer an account of how public discourse, like media and news language, might encode a non-arbitrary relationship between meaning and form. I also outline how this relationship might be learned from language. To do so, I draw on an influential theoretical account of cultural symbols in sociology: linguistic structuralism.

Second, I empirically test whether words' gender suffixed form (e.g., police*woman* vs police officer vs police*man*) corresponds to the gendered ways in which words are used in language. For instance, do words with gender suffixes tend to be used in contexts where gender is more salient? Prior empirical work on gendered language measures meanings of words in human participants directly (what I refer to in this paper as "cognitive meanings"). These meanings could have been learned from a range of possible sources. Therefore, measures of cognitive meanings do not isolate whether the meanings of a word are learned from language or other possible sources (e.g., real word experiences like interactions with police officers). In contrast, I measure the meaning of words, *purely* based on how words are used in natural language (i.e., what I refer to in this paper "linguistic meanings"). To do so, I use a computational text analysis tool: word embeddings.

Word embeddings learn the meanings of words from "reading" an inputted language dataset (e.g., Bojanowski et al. 2017; Mikolov, Sutskever, et al. 2013; Pennington et al. 2014). They thus learn linguistic meaning: the meaning of words based on how words are used in a text dataset alone, rather than based on any non-linguistic experiences. Methodologically, word embedding methods make it possible to empirically test whether the relationship between gendered meaning and gender suffixed form can be learned from how words are used in natural language.[12]

Rather than focusing on all English words with and without gendered suffixes, I focus on a large and important subset: occupational titles. Many occupational titles include gender suffixes (e.g., “mailman,” and “salesman”). The meanings of occupational titles matter because gendered stereotypes of occupations drive the reproduction of gender discrimination and inequality, and gender segregation in the workplace (Ridgeway and Correll 2004). For the same reasons, an abundance of work on gendered language also focuses on occupational titles (Sczesny, Formanowicz, and Moser 2016; Liben, Bigler, and Krogh 2002; Vervecken, Hannover, and Wolter 2013; Vervecken, et al. 2016; Formanowicz, et al. 2013).

I ask three core empirical questions about the relationship between occupational titles' gendered suffixes and their gendered linguistic meanings. First, do occupational titles that have gendered suffixes tend to evoke gendered linguistic meaning that corresponds to their gender suffixes? For example, do titles ending in the suffix “-woman” tend to evoke meanings of femininity, and do titles ending in the suffix “-man” tend to evoke linguistic meanings of masculinity? Second, do titles with gender suffixes evoke *exaggerated* gendered linguistic meanings compared to other words? Third, what gendered linguistic meanings, if any, do titles evoke in the absence of suffixes? I develop three hypotheses for these questions, based on my

[12] As I will explain in the methods section, I specifically use neural word embeddings that process words *and* sub-word information such as affixes, morphemes, and suffixes (Bojanowski et al. 2017).

theoretical account of the relationship between gendered meanings and suffixed form, and prior empirical work on gendered language and cognitive meanings of gender.

This paper begins with a cultural sociological account of symbols and symbolic systems, drawing on a linguistic structuralist framework. Lexical items (e.g., words, phrases, and morphemes) are canonical cases to understand how symbols work. Then, I provide a brief description of gendered language and prior work on the relationship between gender suffixes and gendered cognitive meanings. Leveraging linguistic structuralism, I then offer an account of how the relationship between gendered meaning and form might be encoded in and learned from language. Ultimately, this paper speaks to both cultural sociology and scholarship on gender and gendered language. For cultural sociology, this paper uses gendered language as a case to revisit longstanding questions about cultural symbols. Combined with prior empirical results on gendered language, the empirical results in this paper challenge the dominant view of an arbitrary relationship between meaning, and form. For scholarship on gender and gendered language, this paper complements and extends previous empirical findings and assumptions in experimental psychology on the relationship between gendered meaning and gendered word forms (Sczesny, Formanowicz et al. 2016).

BACKGROUND

Cultural Symbols and Symbolic Systems

Contemporary sociological understandings of cultural symbols are profoundly shaped by a key theoretical movement in linguistics: linguistic structuralism. Here, I provide a theoretical account of symbols, following a linguistic structuralist framework. Following linguistic structuralism, contemporary sociology defines a symbol as the mapping between an external form (e.g., a sound, image, or a series of characters comprising a word) and the meaning evoked in an individual's mind by the form (Lizardo 2016). Together, a set of symbols and their relationships is referred to as a symbolic system. Contemporary sociologists have studied a range of symbols and symbolic systems: from foods (Barthes 1961) to perfumes (Cerulo 2018), the confederate flag (Talbert 2017), condoms (Tavory and Swidler 2009), and even posters (McDonnell 2014). However, across the social sciences, language continues to be the most well studied symbolic system, in which the symbols are words (or multi-word expressions, or morphemes, like suffixes and affixes) and the symbolic system is language.

The relationship between meaning and form.

A core premise of linguistic structuralism is that the relationship between symbols' forms and meanings is largely arbitrary (Craib 1992:137; Joas and Knobl 2009:344).[13] For instance, the five

[13] The premise of arbitrariness is commonly attributed to Saussure. However, the common understanding of arbitrariness in structural linguistics and structuralism is not quite the same arbitrariness as that actually suggested by Saussure (Stoltz 2019). In this paper, I focus on the dominant understanding of arbitrariness: that the relationship between form and meaning is arbitrary.

letters in the word “child” do not resemble what a child looks like, and the word is not pronounced in a way that sounds like a child. More specifically, this premise is that there is an *unpredictable mapping* between the form of signs and their meaning (Dingemanse et al. 2015). If we did not already know the meaning of the word “child,” we would not be able to guess its meaning from seeing or hearing the word alone. In contrast, non-arbitrariness would suggest that aspects of a word’s form can predict the word’s meaning (or, at least, how the word should be used in natural language) (Dingemanse et al. 2015).

This notion of arbitrariness is longstanding and pervasive. By some accounts, arbitrariness is even considered a fundamental and necessary property of language (Hockett 1960). However, recent empirical work in linguistics reveals many correspondences between words’ meanings and their forms (i.e., how words are pronounced and/or spelled) (for a review, see Dingemanse et al. 2015). Two broad types of correspondences are well-documented across languages: iconicity and systematicity.

Iconicity refers to words with forms that resemble sensations associated with the meaning through their form (e.g., Blasi et al. 2016; Perniss et al. 2010). Onomatopoeic words, which were initially highlighted as an unusual outlier to arbitrariness, present just one case of iconicity. I list several more common cases here, drawing on those highlighted by Dingmanse (2012). First, duplication in words’ forms often corresponds with a repetitive meaning. For instance, in Japanese, “goro” refers to a heavy rolling object, while gorogoro refers to multiple heavy rolling objects. In a language spoken in Sierra Leone (Kisi Kisi) hábá refers to humans’ “wobbly, clumsy movement” and hábá-hábá-hábá refers to humans “prolonged, extreme wobbling” movement (Blasi et al. 2016). Second, vowel lengthening corresponds to exaggeration of length, intensity, or duration (e.g., in English, spelling “long” as “looonng” exaggerates how long something is). Third, contrasts in vowel quality can resemble contrasts in magnitude. For instance, in Japanese katakata refers to clattering, while kotokoto also refers to clattering but less noisy. A fourth and well-studied example is that people tend to associate novel shapes to nonsense words, in systematic and consistent ways, based on the sound of the word and roundness (vs angularity) of the shape. For example, participants are more likely to guess that a round shape is matched to a word that requires rounding of the mouth to pronounce (e.g., "bouba is more likely to be matched to a round shape than "kiki") (Köhler 1929; Ramachandran and Hubbard 2001) This is often referred to as the "bouba-kiki" effect. Finally, sign language presents a classic case of iconicity as well: the form of many signs directly represent the motions, shapes, and spatial relations associated with the meaning of the sign (Perniss et al. 2010).

In addition to resembling meaning, words' forms may also be *statistically* related to their meanings (i.e., systematicity). For instance, in many languages, features of words' forms such as vowel quality and syllable duration can distinguish nouns from verbs (Kelly 1992). More generally, across many languages words belonging to different grammatical categories can be distinguished by aspects of words' forms (Monaghan, Christiansen, and Chater 2007). Further, many words are composed of several morphemes; morphological structure is often predictably related to words' meanings. For example, "-er" added to actions (e.g., run → runner, swim → swimmer) often yields a word referring to the noun doing the action. More precisely, systematicity illustrates that the relationship between words' meanings and forms can be *relatively arbitrarily*; we can use other words we already know to predict the meaning of many words. For instance, even if we did not know the word "running," we might be able to infer it is about movement if we knew the word "run" and we knew how "walk" relates to "walking", and how "run" relates to "walk."). In sum, a wide variety of recent work in linguistics, artificial intelligence, and cognitive science highlight that there are both arbitrary and non-arbitrary relationships between form and meaning, across many different languages.

Symbols' meanings are relational.

Another core premise of linguistic structuralism is that symbols' meanings are relational: that is, the meanings of symbols emerge from similarities and differences in how symbols are *used*. For instance, if a word is used around other words about men (e.g., "he" and "man") rather than women, the words' meanings will be more masculine than feminine. Linguistic structuralism specifically identifies two types of relationships: syntagmatic and paradigmatic relationships (Saussure 1916:121–25). Syntagmatic, or associative, relations concern the position of symbols, such as how the order of the words in a sentence combine to create a specific meaning, and how various parts of speech may occur in a sentence. Two symbols are syntagmatically related to the extent that they occur together. For example, the words "apple" and "tree" are often used together and so they are syntagmatically related. Paradigmatic relations concern the substitutability of words; two symbols are paradigmatically related to the extent that they tend to occur in similar contexts. For example, even if "apple" and "pomegranate" rarely co-occur, if these two words are used in similar ways and roughly interchangeable in many contexts, they are paradigmatically related.

In computational linguistics, the notion that meaning is relational resulted in a more specific hypothesis about word meaning: the Distributional hypothesis. This hypothesis suggests that a word's[14] meaning corresponds to the patterned ways in which the word is used in language (Lenci 2018). A more specific and dominant interpretation of this hypothesis is that individuals *learn* the meanings of words (at least in part) from the ways they are used in language (Davis and Yee 2021; Landauer and Dumais 1997; Lenci 2018). In this interpretation, the Distributional hypothesis is a *learning* mechanism for meaning, as well as a definition of words' meanings. Note that this hypothesis, like linguistic structuralism, does not consider extralinguistic factors that can impact our understanding of words (e.g., learning the meaning of the word "nurse" from our interactions with people who are nurses). For clarity, I distinguish meanings in individuals' minds (as "cognitive meaning", which may be learned from a variety of sources) from meaning that may be learned from distributional patterns in language (as "linguistic meaning").

Gendered Language

All human languages encode information about gender to some degree, and they do so in varying ways (Hellinger and Bußmann 2003; Jakiela and Ozier 2018). On the more extreme end, in languages such as Spanish and French, each noun is classified as grammatically feminine or masculine. For example, the word "spoon" in Spanish ("la cuchara") is grammatically feminine, while the word "fork" in Spanish ("el tenedor") is grammatically masculine. Words used to

[14] While the Distributional hypothesis is often discussed in the context of language, it may generalize to other symbolic systems as well (e.g., Arronte Alvarez and Gómez-Martin 2019; Chuan, Agres, and Herremans 2020). In this paper, the symbols I focus on are words (specifically, occupational titles), and the symbolic system is language.

describe each noun, like adjectives, may also be modified to have corresponding masculine or feminine suffixed forms. For example, “the small spoon” translates to Spanish as “la cuchara pequeña” and “the small fork” translates to “el tenedor pequeño.” Across all languages (even those without grammatical gender), words may additionally have gendered *meaning*: for instance, the words “girl,” “girls,” “she,” “woman,” and “women,” are about women. Gendered meaning may or may not correspond to grammatical gender form in these languages (e.g., “girl” is grammatically feminine in Spanish and has feminine meanings).

English does not encode gendered information quite so extensively as do languages like Spanish, French, or German. But English does contain gendered information in a variety of ways. For example, a person’s gender may be specified using terms for social roles (e.g., “grandma,” “grandpa,” “wife,” “husband,” “aunt,” “uncle”), gendered pronouns (e.g., the student liked *her* book), and honorifics (e.g., Mr., Mrs., and Ms.). While English does not categorize all words into grammatically gendered forms, a variety of English words do include gendered suffixes and affixes as part of their form. For instance, the words “princess,” “governess,” and “heiress” include the suffix “-ess;” the words “freshman,” “layman,” and “doorman” include the suffix “-man.” In particular, many occupational titles in English include gender suffixes as part of their form. For this and other reasons, occupational titles are a key site for gendered language in English, as described in the introduction (Formanowicz et al. 2013; Holmes and Sigley 2002; Liben, Bigler, and Krogh 2002; McConnell and Fazio 1996; Sczesny, Formanowicz, and Moser 2016; Vervecken et al. 2016; Vervecken, Hannover, and Wolter 2013).

Gendered language reflects longstanding, societal patterns of inequality around gender (Hellinger and Bußmann 2003; Sczesny et al. 2016). For example, words with male suffixes are often more visible and sometimes used as generics to refer to both men and women (e.g., "freshman"). Similarly, words like "mankind," "to man," "workmanship," "manmade," and "manpower," are used as generics, even though they include male suffixes and affixes. The fact that many words with male forms are used as generics is thought to reflect the "male bias:" an overall androcentric gender system (Lindqvist, Renström, and Gustafsson Sendén 2018; Ridgeway and Correll 2004). More generally, it may reflect that men (rather than women) are the "default" gender. But scholars are not merely concerned that gendered language reflects gender inequalities. The more pressing concern is that gendered language can *influence* our meanings, actions, and perceptions in ways that reinforce gender asymmetries.

Gendered language and its impact.

Abundant work illustrates the wide-ranging impact of gendered language on human cognition (e.g., Boroditsky et al. 2003; Formanowicz et al. 2013; Samuel, Cole, and Eacott 2019; Sczesny et al. 2016).[15] A body of experimental work in cognitive science and linguistics finds that grammatically gendered form is associated with (1) the gendered mental representations evoked in people's minds by the word, and (2) how people understand that concept in non-linguistic scenarios (e.g., Boroditsky et al. 2003; McConnell and Fazio 1996; Phillips and Boroditsky 2003; Sera et al. 1994).

[15] More generally, gendered language presents a case to investigate the relationship between language and thought (i.e., linguistic relativity) (Samuel, Cole, and Eacott 2019; Whorf 1956). Linguistic relativity asks: Does what we say and how we say it affect what we think?

As a key example, Phillips and Boroditsky (2003) found that Spanish and German speakers tended to rate pictures of objects (e.g., toaster, clock, moon, fork, and toothbrush) as more similar to pictures of biological males or females, depending on the object's grammatical gender in their language. Another study found that when a target individual's occupation is described with a male suffix rather than no suffix (e.g., "chairman" vs "chair") participants interpret the target individual's personality as more masculine (McConnell and Fazio 1996). Strikingly, these findings held even when participants *knew* the target individual's gender. This suggests, for example, that when a woman is described as a chairman, she is perceived as more masculine than when she is described as a chair or chairperson. Empirical work in experimental psychology also suggests even when masculine forms of words are used as generics, that these words evoke male exemplars more than female exemplars (Sczesny, Formanowicz, and Moser 2016; Silveira 1980).

Finally, in a timely example, Mecit et. al. (2022) observe that in French and Spanish, the term for the disease resulting from COVID-19 is grammatically feminine ("la COVID-19" in both French and Spanish), while the term for coronavirus itself is grammatically masculine (e.g., "le coronavirus", in French and "el coronavirus" in Spanish). Across a series of studies, they illustrate that grammatical gender of the term (i.e., whether "la COVID-19" or "el/le coronavirus" is used) affects gender stereotypical judgments about the virus. Noting that dangerousness is a stereotypically masculine trait, they further show that people use this stereotypical information to assess the danger of the virus. In sum, it is now well-established that grammatical gendered form can impact the *cognitive* meanings of gender (whether of the word or concept).

A variety of work also illustrates downstream impact of gendered language on outcomes relevant to gender equality (e.g., women's and men's occupational outcomes) (e.g., Bem and Bem 1973; Formanowicz et al. 2013; Pérez and Tavits 2019; Sczesny et al. 2016; Stout and Dasgupta 2011; Vervecken et al. 2013). For example, women perceive themselves as a poorer fit for a job when advertisements or interviews for the job use occupational titles with male gender suffixes for men, compared to titles with female gender suffixes or no gender suffixes (Bem and Bem 1973; Stout and Dasgupta 2011). A job advertisement which uses a job title with male suffixes suggests that the ideal candidate for that position is male, and so women's self-schemas do not align with images evoked by these job titles. This is one example, among many, of how gender asymmetry in our language is one vehicle for gender-based discrimination, as well as stereotyping and inequality (Sczesny, Formanowicz, and Moser 2016). At a more macro level, recent experimental and observational analysis also finds that speaking a language without grammatical gender is related to higher support for efforts to address gender inequality (Pérez and Tavits 2019).

Learning the gendered meanings of words from public discourse.

Why is there a relationship between gendered form and cognitive meanings of words? Where is this relationship learned from? The implicit consensus in research on gendered language is that a cognitive phenomenon known as semantic contagion drives the relationship between gendered language and cognitive meanings of gender (Atagi et al. 2009). Semantic contagion refers to the fact that even if a given word does not evoke some specific information on its own (such as gendered meanings) if it is often surrounded by words that do activate gender information, the corresponding pattern of activation will be reinforced for the given word as well. This is because each time we see and use words, we automatically activate information associated with these words, but are also activating information associated with their context words (Lupyan and Bergen 2016).[16] Across many instances of context words, this suggests how we learn the meaning of words from their contexts.

A well-noted example of semantic contagion in gendered language occurs in languages where *all* nouns have a grammatical ending that is masculine or feminine (such as Italian and Spanish). In these languages, speakers tend to perceive inanimate objects as more masculine or feminine depending on their grammatical endings– even though the objects do not have any inherent gender (Boroditsky, Schmidt, and Phillips 2003). Since these nouns are grammatically gendered, they occur in gendered grammatical constructions. These same grammatical constructions are used for nouns that are semantically and grammatically gendered, such as "woman." For example, in Spanish, "la" is used to refer to words that denote women (e.g., la reina refers to "the queen"), but is also used to refer to objects that are grammatically feminine (e.g., la cuchara refers to "the spoon"). Therefore, words like "la" may impart gendered information to what would seem to be a neutral item, such as a spoon. Thus, even inanimate objects gain gendered meaning from their gendered form and context words.

Crucially, this account of gendered language instantiates the Distributional hypothesis about meaning, described earlier. By this account, we would learn to associate "police officer" with men if police officers are usually described as men or as masculine in conversations, media, news, books, and other language sources we are exposed to. Further, our gendered meanings of

[16] Semantic contagion is often described as, "neurons that fire together, wire together."

words are thought to result from us somehow internalizing the gendered ways in which words are used in public discourse (e.g., Liben et al. 2002). Public discourse can include a wide range of possible sources, such as media, news, fiction, social media, and public data on the web. If natural language is to be a plausible source for learning gendered meanings of words, then we need empirical evidence that gendered forms are used in semantically gendered ways in public discourse, and we need a clearer explanation for the mechanisms involved. Drawing on the Distributional hypothesis and structural linguistics, next I offer a more general account for how the relationship between meaning and form might be encoded in public discourse.

How can distributional patterns contribute to a relationship between form and meaning?

Distributional patterns can encode a relationship between form and meaning in multiple ways. As described earlier, structural linguistics suggests that there are two ways in which words may be related: through syntagmatic and paradigmatic relationships.[17] First, recall that two words are syntagmatically related when they tend to *directly co-occur* in contexts. More generally, if words with particular form (e.g., words with the suffix "-ess") tend to be used next to other words that have gendered meanings, like she and woman (e.g., "she is the hostess," or "that woman is the hostess"), they can gain gendered meanings. Second, recall that two words are paradigmatically

[17] While I distinguish syntagmatic and paradigmatic relations to offer a more specific account of the link between gendered form and meaning, note that I do not empirically distinguish these two types of relations in this paper.

related to the extent they occur in similar contexts, or are substitutable. Therefore, if we use words like warmth or compassion around “policewoman” that we *also* tend to use around “she” more than “he,” the gendered meanings of “she” might spill over to “policewoman” to emphasize the femininity of “policewoman.” More generally, if we use words with particular forms (e.g., words with the suffix “-ess”) in contexts that are similar to words used to describe men or women, then these forms will gain gendered meaning. Most likely, both syntagmatic and paradigmatic relationships drive the relationship between gendered form and meaning. Following this account, I expect that occupational titles with gender suffixes will be used in corresponding gendered contexts in the corpus (H1) and will have stronger meanings of gender than titles without gender suffixes (H2). However, if suffixed form is uninformative in public discourse, we would expect no relationship between titles’ suffixes and their gendered contexts.

This account also explains how titles *without* suffixes can still carry gendered meanings. For example, “police officer” might gain masculine meaning if it tends to be used in the context of words denoting men, like “he” and “man” (i.e., through syntagmatic relations). And, even in the absence of such explicit references to gender, “police officer” might gain masculine meaning if it is described using adjectives that are also often used to describe men (i.e., through paradigmatic relations). These same mechanisms could explain how titles without suffixes could carry feminine meanings as well. Distributional patterns explain *how* titles without suffixes might carry gendered meanings (rather than no gendered meanings). But they do not suggest *which* (if any) gendered meanings titles without suffixes might have.

What meaning (if any) can we expect in the absence of suffixes?

H1 and H2 suggest a correspondence between gendered meaning and gender suffixes. What gendered meaning might titles have, if any, in the absence of gender suffixes? More generally, we constantly try to classify others into social categories (and especially, gender) (Ridgeway 2011:37–43). In the absence of cues (e.g., a description of a person that does not specify gender), we often rely on information such as default assumptions and prototypes for inference (e.g., Reynolds, Garnham, and Oakhill 2006). Our prototypes for a generic person's social groups and characteristics are also often the *dominant* social groups or characteristics (Ridgeway 2011:67–68). For instance, the dominant social characteristics include male, white, heterosexual, and able-bodied, and the prototypical person is also male, white, heterosexual, and able-bodied (e.g., Bailey, LaFrance, and Dovidio 2019; Herz and Johansson 2015; Lindqvist, Renström, and Gustafsson Sendén 2018; Morris 2013; Nosek, Banaji, and Greenwald 2002; Ridgeway 2011; Ridgeway and Correll 2004). Our default assumptions also manifest in language (e.g., Martin and Papadelos 2017). Descriptions that counter our default assumptions (e.g., *male* nurse, or *female* doctor) tend to be specified, while descriptions that do not counter our default assumptions are unspecified and taken for granted (Brekhus 1998; Haspelmath 2006; Zerubavel 2018:32–35).[18]

Given this prior work, I expect that titles without gender suffixes tend to be *described* in ways that match the default assumptions about the gender. More specifically, I expect that most titles without gender suffixes will be used to describe men or used in context of other words used to describe men (rather than women, or no gender), thus having masculine linguistic meanings. Together, H1-H3 suggest a complex relationship between form and meaning: Gender suffixed

[18] Such work refers to this phenomenon as "markedness."

form constrains meaning, but the lack of gender suffixes leaves us with our default meanings of gender, rather than a *lack* of gendered meaning.

Are suffixes meaningful themselves, or are they part of compound words?

In work on gendered language, it also remains unclear to what extent the *suffix itself* matters. It is possible that gender suffixes, on their own, might prime gendered meanings. For instance, the suffix “man” in “policeman” might reinforce the masculinity of the term policeman because the suffix “man” evokes the same words as the standalone word “man.” In this case, we might expect that the gendered meanings of “policeman” compound the gendered meanings of “police officer” and “man.” It is also possible, however, that words are processed as a complete unit, rather than as composed of multiple potential morphological units. The empirical evidence is mixed as to whether humans *see* suffixes separately from the rest of the words, or whether humans process words as complete units, ignoring sub-word information like suffixes (e.g., Leminen et al. 2018; Stites, Federmeier, and Christianson 2016). Given the mixed evidence, I first test whether patterns described in H1-H2 hold up under the modeling assumption that morphological information is explicitly processed when exposed to language. Second, I test these under the modeling assumption that suffixes are not separately processed. Given conflicting empirical evidence on how humans process suffixes in compound words, I make no hypotheses for how H1 and H2 hold up under one model versus another.

METHODS

Approach

To test the hypotheses described in the previous section, I follow three core steps. First, I collect and clean a news corpus and a dataset of occupational titles. Second, I model the distributional meanings of all words in this news corpus, using the machine-learning tool neural word embeddings. I do this step in two *types* of models: first, one which processes sub-word information such as suffixes, affixes, and morphemes (i.e., FastText (Bojanowski et al. 2017)), and second, one which processes only word-level information (i.e., word2vec (Mikolov, Sutskever, et al. 2013)). To be clear, word2vec is *not* given sub-word information (such as suffixes) separately from whole words (i.e., it does not learn suffixes independently of words with suffixes). Thus any relations found between gendered meanings cannot be explained by suffixes' meanings alone. Third, I measure the gendered meanings of occupational titles learned by this model, and then I perform statistical analyses to compare gendered meanings by suffixed form. I additionally replicate these findings using a FastText word embedding trained on a broader set of public discourse. I will describe these steps in more detail next.

News Text Data

News discourse is one of many important contexts in which to study language about gender (Sendén, Lindholm, and Sikström 2014; Shor, et al. 2015). News reporting is a widely circulated and authoritative source of social information. I also argue that, in contrast with sources of public discourse such as social media, news is a more authoritative source for *how to talk about* social information, such as gender. It may thus be an influential source for learning linguistic meanings of social roles, including occupational titles, from language. The *New York Times*, in particular, is a widely circulated and leading source for news ("Top 10 U.S. Daily Newspapers" 2019). In my main analyses, I model the language of news discourse in the *New York Times* Annotated Corpus. Specifically, I use the articles published between 2002 and 2007 in this corpus; this uses the most recent years of the corpus and years during which there were no changes to the *New York Times* style guide. These data include 499,864 *New York Times* articles.

In addition, I replicated all analyses using a public FastText model pre-trained on English language news data, Wikipedia meta-pages, and web-crawled text (Mikolov et al. 2018). More specifically, the training data included English Wikipedia from June 2017 (9 billion words) and all news datasets from statmt.org from 2007 to 2016 (4.2 billion words) as well as web data crawled in February 2007 (i.e., the University of Maryland, Baltimore County (UMBC) web base corpus) comprising 3.2 billion words (Han et al. 2013). A disadvantage of using pre-trained models, including this one, is that their training data and construction (e.g., hyperparameters and data cleaning choices) are not transparent. Further, it is not possible to process and clean the raw text data in custom ways (e.g., to match the choices I made for cleaning the *NYT*, or to limit to just *news*). However, the advantage is that pre-trained models are trained on a very large variety of data (which is impractical to impossible to access directly). Therefore, replicating results using a pre-trained model, such as this one, offers an efficient way to test the generalizability of results beyond the *New York Times* corpus used in main analyses.

Identifying Occupational Titles and Manually Annotating Titles for Gender Suffixes

My goal in identifying a list of occupational titles was to collect as many official and colloquial ways to refer to occupations as possible, limiting my list to singular forms of titles. I began with occupations in the 2010 Census (Statistics 2016). I added additional titles while conducting a literature review using the search terms "gender-fair language," and "gender-fair language and occupations." This process yielded a candidate list of 877 possible occupational titles. Finally, I excluded titles that are included in the text data fewer than 20 times, since word embeddings need multiple instances of a word to learn its representation well, which left the final list of occupational titles used in analyses (N=460). It may be unsurprising that a limited number of occupations is represented in the news: the 2010 Census list contains many very specific occupations that are rarely used in everyday language (e.g., stucco mason and plasterer).

For occupational titles which included more than one word (e.g., police officer), in the text data, I used regular expressions in Python to replace the title with a version using underscores rather than spaces between the words (e.g., I replaced "police officer" with "police_officer"). This data cleaning step ensured that these multi-word titles would be represented as single word-vectors, rather than as two separate word-vectors.

With this final list of 460 occupations, I then manually labeled occupations as having masculine suffixes, feminine suffixes, or no gender suffixes. I define these gender suffixes based on a literature review (e.g., Bem and Bem 1973; He 2010; Hellinger and Bußmann 2003; Holmes 2001; Holmes and Sigley 2002; Liben et al. 2002), recognizing that there is some ambiguity around some gender suffixes.[19] More specifically, I labeled any titles with the suffix "-man,"

[19] I also labeled "midwife" as having a feminine suffix, expecting that the suffix "-wife," functions similarly to a gender suffix (as in "housewife"). I also labeled "housemaid" as having a

"-boy," or "-master" as having a masculine suffix. I labeled any titles with the suffix "-woman", "-ess", "-lady", "-maid", "-ienne," "-girl", or "-wife" as having a feminine suffix. Some examples of occupations that I labeled as having gender suffixes, following this scheme include: policeman, garbageman, tv_repairman, cleaning_woman, waitress, and hostess. Among the 460 occupational titles investigated in this study, this manual annotation process yielded 55 titles with gender suffixes (24 with feminine suffixes and 31 with masculine suffixes).

Modeling Gendered Linguistic Meanings of Occupational Titles with Word Embeddings

To model the linguistic meanings of occupational titles, based on how they are used in natural language, I use the computational text analysis tool word embeddings. Word embeddings model the meaning of each term in a corpus as a vector space. Specifically, given a large corpus, they learn from the data to represent each vocabulary word's meaning as an *N*-dimensional vector (a list of N numbers). Just as a 2-dimensional vector locates a position in a plane (a space that is defined by two dimensions, the "X" axis, and "Y" axis), an *N*-dimensional vector locates a position in an N-dimensional space. Thus, each word's vector may be thought of as locating a word in N-dimensional space. This space is also called a word embedding or semantic space

gender suffix, since "maid" is gender-specific in its definition in Merriam Webster, much like "woman" and "wife." Given ambiguity around the gender specificity of "-er" and –"or" (e.g., waiter, actor, banker, teacher) (Liben, Bigler, and Krogh 2002), I do not label titles with these endings as having gender suffixes.

since it encompasses all the meanings of the vocabulary words. The dimensionality of the space (and thus vectors) is set by the researcher, often between 100-500 dimensions (Rong 2014).

While many word embedding architectures (e.g., GloVe and Word2Vec) only process incoming text as whole *words,* FastText processes incoming text data at the character level, as described next. This means that FastText learns morphological or "sub-word" information for words, including suffixes and affixes. More generally, this means that FastText does not just learn to represent words as vectors, but all morphemes as vectors.

To learn to represent morphemes as vectors (morpheme-vectors) from the raw text data alone, FastText models may use the algorithm "Continuous Bag of Words," or, CBOW.[20] With CBOW, FastText learns morpheme-vectors by iteratively predicting target words from their various contexts (surrounding morphemes) in the text data and updating morpheme-vectors to minimize prediction errors on this task. For example, consider a hypothetical context for the word "policewomen:" "There were many policewomen at the crime scene." FastText is given just the context words around "policewomen" in this phrase as a series of morphemes, e.g., "there," "were," "many," "at," "the," "crime," "scene," and predicts the word with the highest probability of being the missing word – equivalently, the word which is closest in space (more similar) to the context words. Each word is inputted as a sum of their character n-grams- for policeman, for example, with n=3, this is: <po pol oli lic ice cem ema an> policeman. A word's

[20] One of two possible algorithms may be used (CBOW or SkipGram), both are similar algorithms and yield similar resulting word-vectors. In contrast with CBOW, the task in SkipGram is to guess context words/morphemes from a target word/morpheme.

representation can be accessed in two ways: as the vector for the word or as the sum of the vector representation of all the character n-grams.

Initially, before any learning, each morpheme-vector is random, and so the predicted morphemes are often incorrect compared to what word is in the phrase. Across many instances of predicting morphemes from their contexts, FastText adjusts these initially random morpheme -vectors to improve prediction accuracy. Specifically, it shifts morphemes' vectors closer together (i.e., makes their representations more similar) to improve prediction accuracy. Thus, these adjustments iteratively improve how well the morpheme-vectors capture the meanings of words; such as that "crime" and "scene" probably relate to "policewomen." Conceptually, if the morpheme-vectors can perform this task, they must have encoded some aspects of meaning.

In a trained FastText model, vocabulary words and morphemes which share more aspects of meaning have similar vectors. Since vectors locate positions in space, similarity and distance are interchangeable; words with more similar vector representations are also closer in space. This similarity, or distance, may be measured with cosine similarity. Cosine similarity between two FastText word vectors ranges from -1 (exactly the opposite), to 0 (not at all the same), to 1 (exactly the same).

FastText learns meanings merely from regularities in inputted text data. Thus, FastText has no reference to the real world as it learns and begins as a blank slate before encountering training data, unlike a human participant. In other words, it allows us to see the meanings learned from language alone, rather than from language and other sources. In line with a linguistic structuralist account and the Distributional hypothesis, FastText learns these words by predicting words from their linguistic contexts (Firth 1957), and models words as part of a meaning *system* where words are defined by their relations (geometric relationships) to other words.

Implementation of FastText.

To train FastText models, I use the Python package Gensim (Rehurek and Sojka 2010). I use CBOW and set the context window to be 10 words before and after each target word and the dimensionality of the word-vectors at 300. Word embedding models can be unstable across various training instances (Antoniak and Mimno 2018). Therefore, I train 15 FastText models on the corpus and use average gendered meanings across these 15 models.

It is important to evaluate word embedding quality, however, this is also an open and ongoing area of research (e.g., Faruqui et al. 2016; Gladkova and Drozd 2016; Wang et al. 2019). Two common (if imperfect) metrics for evaluating the quality of trained word embedding models include the Google Analogy Test and a comparison with human-rated similarities. The Google Analogy Test includes nearly 20,000 analogies across various sections (e.g., family, currencies, tense) for a trained model to complete (Mikolov, Sutskever, et al. 2013, Mikolov, Chen, et al. 2013, Pennington, Socher, and Manning 2014). A standard human-rated similarities test set is WordSim353, which includes 353 similarities rated by at least 13 individuals (Finkelstein et al. 2002). My 15 FastText models correctly complete an average of 47% (SD= 0.108%) of analogies on the Google Analogy Test across all sections. Among the family section, which is most relevant to the domain of occupational titles and gender, the models complete an average of 61% of analogies (SD= 0.91%). The 15 models also have an average Spearman correlation of 0.51 (SD=.0030; each p<.0001) with human-rated similarities on the WordSim Test. This performance is somewhat lower compared to published, comparable word embedding models and FastText models (Mikolov, Sutskever, et al. 2013, Mikolov, Chen, et al. 2013, Pennington, Socher, and Manning 2014). However, as mentioned earlier, results replicate on both a word2vec model and a pre-trained FastText model trained on a far larger dataset. Further, detailed scoring procedures are sometimes inconsistent across papers; thus this comparison should be interpreted loosely.

Previous work using word embedding to study gendered linguistic meaning.

A robust body of work demonstrates that word embeddings trained with these approaches provide meaningful representations of text data (e.g., Bolukbasi et al. 2016; Caliskan et al. 2017; Charlesworth et al. 2021; Garg et al. 2018; Grand et al. 2018; Joseph and Morgan 2020; Kozlowski et al. 2019). For example, trained word embeddings can solve a variety of syntactic and semantic analogy tasks and can evaluate similarities between words (e.g., Bojanowski et al. 2017; Mikolov, Sutskever, et al. 2013; Pennington et al. 2014). Further, a variety of work demonstrates that the semantic information captured by embeddings trained on public discourse (e.g., news, web data, media, fiction, or Wikipedia) strongly correlates to survey respondents' meanings (e.g., Bolukbasi et al. 2016; Caliskan et al. 2017; Charlesworth et al. 2021; Garg et al. 2018; Grand et al. 2018; Joseph and Morgan 2020; Kozlowski et al. 2019). For instance, Caliskan et. al., (2017) found that gendered semantic information about names and academic subjects in word-embeddings correlated to human participants' meanings of these names and subjects (measured using the implicit association test).

Several prior studies have specifically used word embeddings to examine the gendered meanings of occupations in cultural texts (Bolukbasi et al. 2016; Caliskan et al. 2017; Charlesworth et al. 2021; Garg et al. 2018; Kozlowski et al. 2019). Such work provides evidence for a strong correlation between how occupational titles are gendered in public discourse with 1) percentages of women and men in these occupational roles and 2) perceptions of how occupations are gendered. For example, Garg et al showed how occupations' gendered meanings in models trained on Google Literature Fiction from 1900-2007 change across time in ways consistent with census data on the percentages of women and men in these occupations (2018). Using an even wider range of corpora, Charlesworth et. al. found gendered meanings of over 300 occupations in word embeddings moderately correlated with percentages of women and men in these occupations (2021). Further, Caliskan et. al. showed that the way that occupations are gendered in Google News, and in the Common Crawl Corpus[21], strongly correlates to census data on the percentages of women and men in these occupations (2017). Together this body of work provides robust evidence that word embedding methods can capture how words (including occupation titles, specifically) are described in gendered ways in cultural texts.

Measuring gender linguistic meanings with word embeddings.

To measure gendered linguistic meanings of words in text, I follow validated methods used in prior work (Caliskan et al. 2017; Garg et al. 2018; Grand et al. 2018; Kozlowski et al. 2019). First, I begin with a set of 20 word-vectors representing men (e.g., "he") and 19 word-vectors representing women (e.g., "she"), drawing these from words identified by Garg et al (2018). I average word-vectors about women, and average the ones about men, and then subtract these two averages. This yields a line corresponding to gendered meaning, which ranges from a point in semantic space "about women, and not men" to a point "about men, and not women." This step

[21] This is a very large corpus containing 8 years of crawled web page data.

builds on the finding that word embeddings not only learn the meaning of words as points in space but encode relational concepts, such as gender, as lines between these points (Arseniev-Koehler and Foster 2020; Grand et al. 2018; Kozlowski et al. 2019; Larsen et al. 2015).

Second, to measure the gendered meaning of any given word, like "hostess," I compute the cosine similarity between the word-vector for "hostess" and the gender dimension. Conceptually, this corresponds to how "gender" makes up the meaning (or, location) of the word-vector for "hostess" in the semantic space. This cosine similarity yields a positive or negative scalar hypothetically ranging from -1.0 to 1.0[22], where a more positive scalar suggests that "hostess" is more feminine, and a more negative scalar suggests that "hostess" is more masculine. For interpretability, in the results, I transform this scalar into a standardized score. Specifically, I standardize a given cosine similarity against the cosine similarities between *all* words in the corpus and this gender direction, yielding a final gendered "score" corresponding to the number of standard deviations more about women (or men) the word is compared to all other words in the corpus (Boutyline et al. 2020).

Hypothetically, the cosine similarity (between a word and the gender dimension) of 0 corresponds to no gendered meaning. However, in practice word-vectors are noisy. Thus, to allow titles to be neutral in gendered meaning, I compare them to a null distribution of words we might expect to lack gendered semantic information: stopwords. Stopwords are words that are very frequent and thought to be meaningless, such as "the" and "a" (Bird, Klein, and Lope 2009). I define stopwords using the NLTK toolkit in Python (excluding stopwords related to gender,

[22] I normalize word-vectors and gender-directions to be unit-length before using them for analyses.

such as gender pronouns). A total of 124 stopwords are in all 15 FastText models. I define an occupational term as carrying gendered meaning if it is significantly more gendered than are stopwords; otherwise, I manually set the gendered meaning to 0 (neutral). Thus, my final estimates for gendered meaning are masculine, feminine, or not gendered. In analyses, I report these categorically, or as the gender score (i.e., standardized cosine similarity to gender). The absolute value of the scalar corresponds to gender meaning irrespective of which gender the word is closest to (masculine or feminine).

There is an important caveat to this measure of gendered meaning. The goal of this measure of gendered meaning is not to capture any true scope of gender *identity*, but rather a pervasive, hegemonic way that gender is represented in popular culture and language in particular – as a dichotomy between men and women (Kachel, Steffens, and Niedlich 2016).[23]

Statistical Analyses

To compare the gendered linguistic meanings of occupational titles based on their gender suffixes (while making limited assumptions about distributions of gendered cosine similarities) I use non-parametric statistics as implemented in R. For comparison between groups, I use Mann-

[23] As this measure ranges continuously from more masculine to more feminine along a line, it also implies that a word which is more feminine, is also less masculine (and vice versa). Previous work by Grand et. al. (2018) additionally finds that measuring gender in embeddings as mutually exclusive and lying on a dimension, corresponds more closely to popular representations of gender compared to measuring femininity and masculinity as two independent clusters of meaning.

Whitney tests. For confidence intervals, I used rank-based measures (Bauer 1972; *R: A Language and Environment for Statistical Computing* 2018).

RESULTS

Among the 460 occupational titles examined in this study, most (N=343) were associated with masculinity, while 86 were associated with femininity, and 31 did not carry specifically masculine or feminine gendered linguistic meaning. For brevity, in the results section, I refer to gendered linguistic meaning as simply gendered meaning. The fact that occupations are strongly stereotyped and segregated by gender also makes sense given that women's and men's social roles are historically segregated by gender (Eagly and Steffen 1984; Ridgeway 2011). The mean gendered score was -0.33 (range -3.36 to 3.51). This score means that on average, occupational titles were 0.33 standard deviations more masculine than the average word in the corpus. Recall that higher, positive scores indicate more feminine meanings, while lower, negative scores indicate more masculine meanings, and that this score is normalized against gender associations of all vocabulary words in the model. The fact that occupations tend to have masculine meanings makes sense given that breadwinning is traditionally a masculine role (while caretaking is a traditionally feminine role) (Nosek et al. 2002; Ridgeway 2011).

The 10 most feminine occupations are saleswoman, chairwoman, actress, forewoman, anchorwoman, waitress, businesswoman, cleaning woman, hostess, and registered nurse, which have gendered scores ranging from 2.59 to 3.51. Thus, for instance, "registered_nurse" is 2.59 standard deviations more *feminine* than the average word in the corpus. The 10 most masculine occupations are foreman, rifleman, crewman, sportsman, tradesman, fireman, barber, gunner, salesman, and statesman. These occupations have gender scores ranging from -2.49 to -3.36. For instance, “statesman” is 2.49 standard deviations more masculine than the average word in the corpus. See figure 1 for the gender meanings of a random sample of 50 occupational titles.

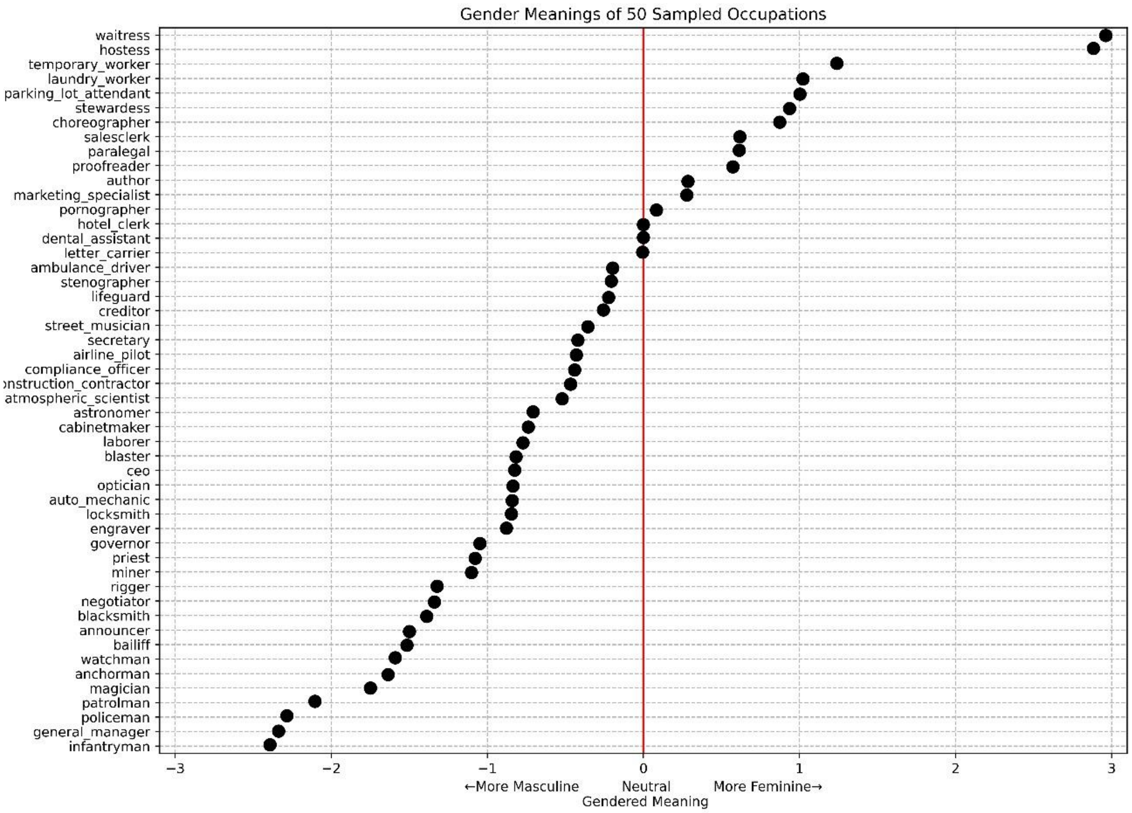

Figure 2. Gendered meanings for a random sample of 50 occupational titles, where meaning is based on use in the New York Times 2002-2007. The gendered meaning of an occupational title is standardized based on the gendered meanings of all words in the corpus.

Occupational Titles’ Gendered Meanings Correspond to their Gender Suffixes

Among the 55 occupations with gender suffixes, all but one have gendered meanings that match their suffixed form. Occupations with feminine suffixes are significantly more feminine than occupations without feminine suffixes (p<.001). The 95% confidence interval for the difference is (2.15, 3.02). Similarly, occupations with masculine suffixes are significantly more feminine than occupations without masculine suffixes (p<.001). The 95% confidence interval for the difference is (-2.10, -1.48).

Among the occupations with feminine suffixes, the median gendered meaning is 2.11 (Mean= 2.11, SD=0.97). Among the occupations with masculine suffixes, the median gendered meaning is -2.12 (Mean= -2.05, SD=0.75). These results suggest that masculine-suffixed words carry exaggerated masculinity, and feminine words carry exaggerated femininity compared to the average word in the corpus. Put another way, gender-suffixed form corresponds to gendered meaning but is associated with very exaggerated gender meaning, both for masculinity and femininity. Thus, these results corroborate hypotheses one and two.

Table 1. Gender Meanings of 460 Occupational Titles

	Feminine Meanings (N)	No Gender Meaning (N)	Masculine Meaning (N)	Total (N)
No Gendered Suffix	63	30	312	405
Female Suffix	23	1	0	24
Male Suffix	0	0	31	31
Total	86	31	343	460

Note: Gendered meaning based on how occupational titles are used in The New York Times 2002-2007, measured using FastText.

Occupational Titles with Suffixes Carry More Gendered Meanings than Titles without Gender Suffixes

Titles with gender suffixes tend to carry exaggerated meanings of gender (whether feminine or masculine) compared to titles without gender suffixes (p<.0001). To get a sense of the extent of this difference, consider the median gendered meaning of titles with suffixes (whether feminine or masculine meanings) is over three times the amount for titles without suffixes (among titles with gender suffixes: median=2.12, Mean=2.072, SD=0.59; among titles without gender suffixes: median 0.64, Mean= 0.73, SD=0.85). Figure 2 shows the distributions of gender meanings of titles, by their gender suffixes. These findings support that gender suffixed form is associated with more salient gendered meaning.

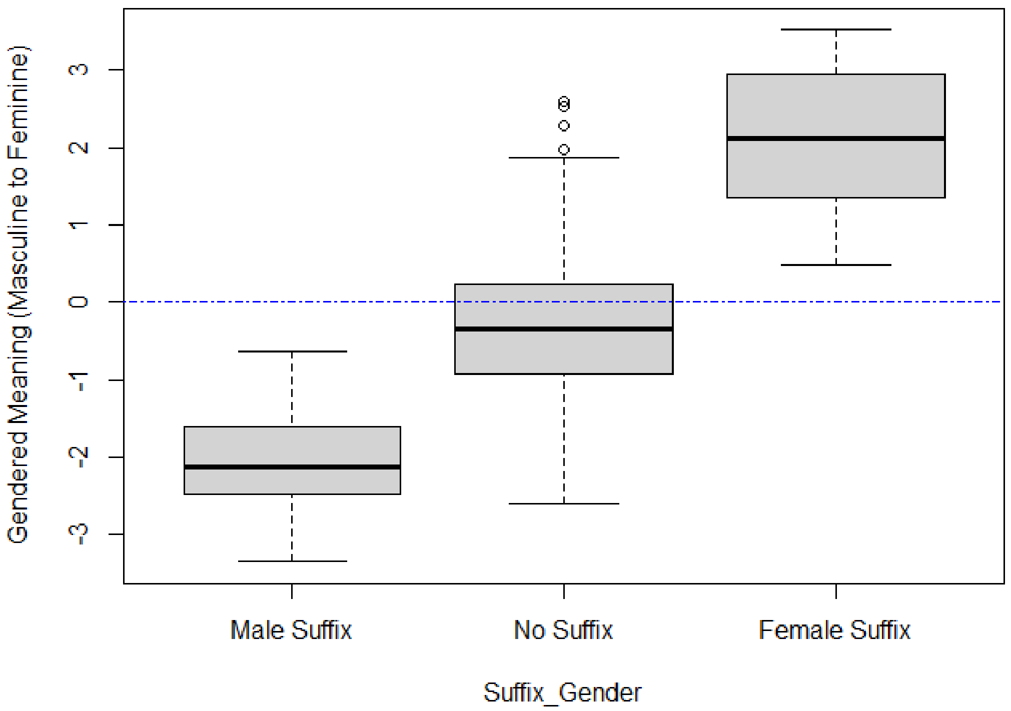

Figure 3. Gendered meanings of 460 occupational titles with and without gender suffixes, using FastText-based measures of gendered meaning trained on New York Times data. Gendered meaning (y-axis) is the number of standard deviations more feminine (+) or more masculine (-) compared to the average word in the corpus.

Even Occupational Titles without Suffixes Carry Gendered Meanings

Even among occupational titles without gender suffixes, most (81%, N=375) carry gendered meanings. Among these, most (83%, N=312) have masculine meanings. The mean gendering score was -0.34, suggesting that even titles without gender suffixes tended to have about a third of a standard deviation more masculine meanings compared to the average word in the corpus. Sample occupational titles which carried masculine meanings include “doctor,” “lawyer” and “musician,” no gendered meanings include “physician_assistant,” “urban_planner,” and “salesperson.” Those with feminine meanings include “sales representative,” “paralegal”, “dietician.”

These Results Hold even if Occupational Titles are Processed as Compound Words

One possibility described earlier, is that suffixes inside occupational titles are processed as separate words. For example, an individual may see the word “policeman” as including the word “man.” Therefore, I also test whether prior results hold when titles with suffixes are seen as compound words (e.g., “policeman” is seen as a single entity). As described in the Analyses, I repeat analyses for questions 1-3 using word embedding models trained on text data where occupational titles are seen as compound words (i.e., using word2vec). Figure 3 illustrates the gendered meaning of occupational titles in this alternate model.

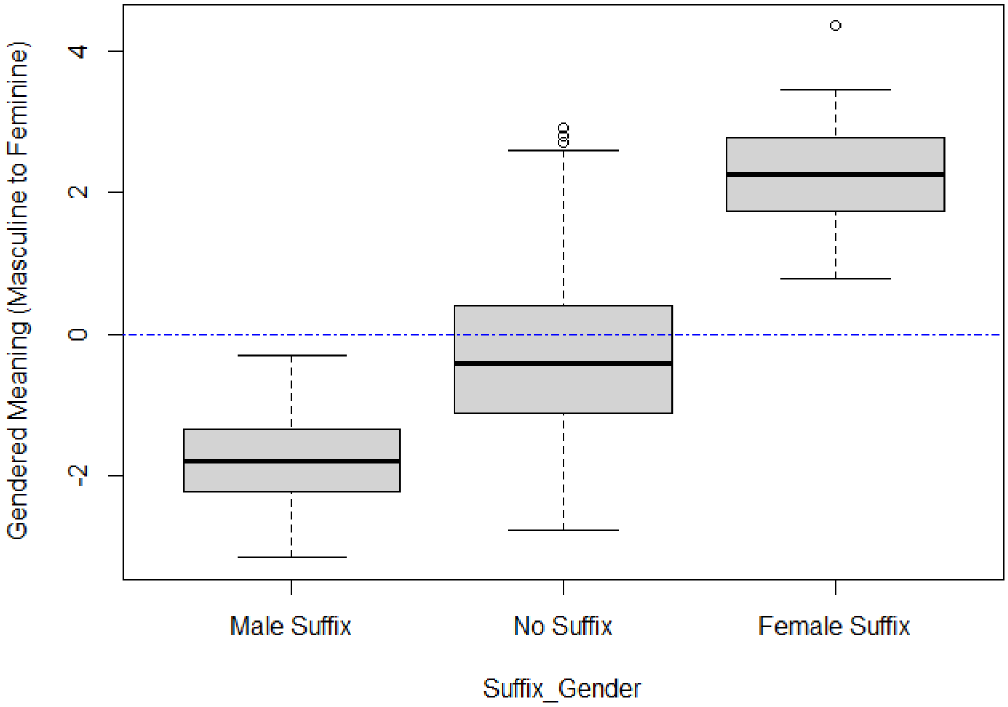

Figure 4. Gendered meanings of 460 occupational titles with and without gender suffixes, using word2vec based measures of gendered meaning trained on New York Times data. Gendered meaning (y-axis) is the number of standard deviations more feminine (+) or more masculine (-) compared to the average word in the corpus.

All conclusions remain in this word embedding model which does not account for morphological form. Specifically, all words with male suffixes have masculine meanings (mean = -1.71), and all words with female suffixes have feminine meanings (mean = 2.29). Titles without suffixes still have masculine meanings on average (mean= -0.29). Titles with gender suffixes have significantly more gendered meanings (whether masculine or feminine) than do titles without suffixes (p<.0001). And finally, titles with suffixes have significantly more gendered meaning (whether feminine or masculine) than do titles without suffixes (p<.0001). The fact that results hold while using this revised model suggests that results are not just driven by the *suffix* itself. This correspondence between meaning and form occurs because titles with suffixes tend to be *used* in context words that are more gendered, compared to titles without suffixes.

To provide more qualitative intuition around the role of context words in differentiating meanings of titles by their suffix, focus on a set of titles that provide a version with no gender suffixes (police officer), with feminine suffixes (policewoman) and masculine suffixes (policeman). I identify the overlapping words among the top 25 closest words to each of these three words, and in Table 2 below I list these overlapping words. As illustrated in Table 2, the words which are closest to “policewoman”, but *not* close to “policeman” or “police officer”, include sexualized and feminized words such as “stripper,” “sexpot,” and “bride-to-be.” Notably, no pronouns are included in this list, and while “policeman” and “policewoman” include many words that explicitly denote gender (e.g., “frenchwoman”, “woman”, “schoolgirl”, and “militiaman”), words closest to “police officer” do not.

Table 2. Overlaps and differences among the top 50 closest words to: Police officer, Policewoman, and Policeman

Set	**Terms**
Terms Closest to Police Officer (and not policeman or policewoman)	assailant, bouncer, detectives, dispatcher, driver, firefighter,landscaper, lieutenant, off-duty, officer, officers, patrol_officer, police, police_officers, priest, vasquez
Terms Closest to Policeman (and not police officer or policewoman)	attacker, crewman, demonstrator, haji, hotel_clerk, interrogator, looter, militiaman, paratrooper, passer-by, policemen, shopkeeper, tribesman, villager, warlord
Terms Closest to Policewoman (and not policeman or police officer)	bank_manager, bride-to-be, brunette, cad, divorcée, frenchwoman, girl, hairdresser, hoodlum, mobster, nun, playmate, private_detective, psychopath, quadriplegic,

	redhead, schoolgirl, schoolmate, sexpot, socialite, sociopath, stripper, transvestite, widower, woman
Terms close to police officer AND policeman AND policewoman	bodyguard, bystander, cabdriver, co-worker, construction_worker, cop, man, medic, patrolman, postal_worker, schoolteacher, security_guard, soldier, state_trooper, taxi_driver, truck_driver

Note: For clarity, these terms come from just the first of the 15 word2vec models.

DISCUSSION

An abundance of empirical work in psychology and linguistics documents a correspondence between grammatically gendered form and gendered meanings of words *in personal culture* (e.g., using surveys and experiments with human subjects). It is often implicitly assumed that a source for learning this correspondence between gendered meaning and form is *the relational ways in which words are used in public culture* (Atagi et al. 2009). This paper unpacks this assumption and tests it, focusing on the way English occupational titles with and without gender suffixes are used in news and other cultural discourse. Results demonstrated that the relationship between gendered meaning and suffixed form (among occupational titles in English) exists in news and a range of other public discourse. Further, results suggested that this relationship can be *learned* from the distributional patterns of words in news discourse.

More generally, results in this paper reveal the complex relationship between form and linguistic meaning in the case of gendered language. Titles with feminine suffixes tended to be used in context of words about women, and titles with masculine suffixes tended to be used in the context of words about men. Titles with gendered suffixes also tended to have more potent gendered meanings compared to titles without gendered suffixes. But it is not necessarily *the* *form* itself that evokes meaning. Word embeddings trained on data blind to morphological form still learned a relationship between suffixed form and gendered linguistic meaning. This suggests that words with gender-suffixed forms are not more gendered because suffixes are processed as separate words (e.g., "man" and "congressman"), but because words with suffixed forms tend to be used in more strongly gendered contexts. These contexts could be words that explicitly denote gender, like "man," "men", and "he," but they can also be adjectives and other types of words.

To provide intuition around the role of context words, in the results I listed examples of words that are closely related to policeman, police officer, and policewoman, as well as words that were uniquely related to each word. Notably, words that were used similarly to policewoman, but not police officer or policeman, included numerous sexualizing words and stigmatized identities, such as "sexpot" and "sociopath." The importance of context words may be unsurprising from a structuralist point of view, which envisions meaning as a *system*. In this view, gendered meaning is distributed throughout our language rather than reliant on any specific linguistic features like gender pronouns or suffixes.

The importance of linguistic context explains another common finding in work on gendered language: that even titles with no gender suffixes tend to have gendered meaning. In my results, these titles, too, tend to be used in the context of gendered discussions – primarily, discussions about men. These results underscore that the *lack* of gender-suffixed form does not imply the lack of gendered meaning. Marking (e.g., through a gender suffix) brings visibility to an attribute, and can reorient our attention away from our default meanings (Zerubavel 1993; Zerubavel 1996). In the case of occupations, since the prototypical worker is masculine (Nosek et al. 2002), it makes sense that *unmarked* occupations (i.e., those without gender suffixes) tend to have meanings of masculinity, rather than femininity.

Results in this paper make an important contribution to cultural sociological understandings of symbols. Theoretical accounts of cultural symbols in sociology are rooted in a structuralist vision of meaning. Core premises of linguistic structuralism include that a symbol's form is analytically separable from meaning, that there is an arbitrary relationship between form and meaning, and that a symbol's meaning may be defined, at least in part, by the ways the symbols are *used* in relation to other words. Results in this paper demonstrated a systematic relationship between gendered meaning and suffixed form in public culture, in the case of occupational titles. This relationship occurs because titles tend to be used in systematically different ways, depending on their suffixes. This paper also highlighted a range of other recent empirical work in cognitive science and linguistics which similarly suggests that the distinction between meaning and form is more complex than theoretical accounts in cultural sociology suggest.

These results also contribute to work on gender and gendered language. First, a core topic in gender scholarship is gender differentiation: when, how, and why do we differentiate genders and gender-related qualities? This paper highlights the potential role of language in gender differentiation. To the extent that we *learn* meanings from language, language is one mechanism by which we constantly categorize, and differentiate, men and women. Gender is categorized and differentiated through the patterned ways we use suffixed forms, pronouns, and more subtly gendered context words. Second, occupations are especially well-known to be persistently and dramatically segregated along gender lines (Cohen 2013). This segregation contributes to gender inequality in that men tend to hold higher prestige (and higher-paying) positions compared to women (Cohen 2013). Gendered stereotypes of occupations drive the reproduction of gender discrimination and inequality, and gender segregation in the workplace (Ridgeway and Correll

2004). This paper contributes to a large body of recent work on how stereotypes and sex segregation of occupations are reflected in public discourse (e.g., Bolukbasi et al. 2016; Caliskan et al. 2017; Garg et al. 2018). This study specifically illustrated that occupations tend to have masculine linguistic meanings and that titles with gender suffixes have corresponding and exaggerated gendered linguistic meanings in a variety of public discourses.

Limitations and Future Research

This study empirically examined the linguistic meanings of occupational titles in the news, specifically the *New York Times* Annotated Corpus 2002-2007. As described in the methods, I additionally replicated all results in a public FastText model which was trained on a far larger dataset of news (across a wider timescale), web-crawled data, and Wikipedia data. The replicability of my results in this pre-trained model suggests that findings on the relationship between gender suffix and the gendered meaning of occupational titles generalize to natural language in a range of news and information sources, and to a range of time periods. Future work might test whether these patterns hold across even wider linguistic contexts and time periods. Specifically, it might test whether patterns hold across corpora that better reflect where individuals might learn the meanings of words, such as fiction, social media, and conversations. One potential example of a possible range of text datasets is included in Charlesworth et al. (2021).

Future work might also more comprehensively test *which* context words or types of contexts drive the relationship between gender suffixed form and linguistic meaning. For example, to what *extent* is this relationship driven by pronouns, or by adjectives, or verbs, or specific sets of adjectives and verbs? One potential tactic to test this is to train word embedding models on text with these different types of words or contexts *removed* and then measure the impact on the amount of gender meaning and the relationship between gendered meaning and gendered form.

While this study focused on the relationship between meaning and form, gendered language also offers many potential insights into other aspects of cultural symbols. Future sociological work on cultural symbols could also continue to theorize and investigate the implications of recent advances in psychological and linguistic work on the linguistic relativity of gendered language (e.g., Boroditsky et al. 2003; Deutscher 2010; Phillips and Boroditsky 2003; Samuel et al. 2019). More broadly, work on linguistic relativity offers an important and well-studied site for sociologists to expand theoretical work on how symbols can impact what we think and what we do.

Finally, this paper intentionally focused on gendered meanings learned from language alone. However, future work on gendered language might also explicitly examine the extent to which and how we learn gendered meanings from linguistic *and* non-linguistic sources. For example, we might learn to associate the word "police officer" with men because we *interact* with more police officers who are men. Indeed, just 13.7% of police officers were women in 2007 (*2017 American Community Survey* 2017). It is also possible that linguistic and non-linguistic sources interact or offer the same information. Indeed, given that more police officers are men, reporting on police officers is more likely to describe police officers as men rather than women, holding all else equal. Previous work demonstrates that occupations' gendered meanings in news (using word embeddings trained on news data, including the *New York Times*) correspond tightly (but not perfectly[24]) to the proportion of women vs men in these occupations (Garg et al. 2018). One interesting direction for future work is identifying the *extent* to which gendered meanings are learned (directly and indirectly) from various linguistic and non-linguistic sources.

Implications

This paper has several implications for popular and scholarly debates about gendered language. To reduce gender biases in language, the Gender-Fair Language movement proposes linguistic interventions, such as to either mark words equally for gender (e.g., "policewoman/policeman") or to unmark words for gender (e.g., "police-officer"), and alter how we use pronouns. The results presented in this paper suggest that to address gender asymmetries in language, we must recognize that gendered meaning is not necessarily isolated to the word itself (e.g., whether the word includes a suffix or not), but the *way* the word is used relative to other words. Similarly, recent work illustrates how job advertisements for female-dominated professions contain more descriptions of particularly feminine soft skills such as empathy, respect, sensitivity and

[24] Representations of occupations in language also can diverge from real life patterns. For example, Garg et al (2018) also found that occupations which are held by an *even* number of women and men tend to be used in contexts about *men* when described in news. This result echoes the finding in this paper that occupational titles tend to be masculine (including occupational titles that have no gender suffixes).

dedication, and less description of stereotypically masculine soft skills like marketing skills, ability to win new business and to lead projects (Calanca et al. 2019). Thus, to be truly gender-fair, we not only need to intervene in the use of gender pronouns and suffixes, but also avoid using gender-loaded contexts and descriptions (e.g., gendered uses of the word "assertive" and "compassionate."). Successfully intervening into gendered language will require us to not just address which words we use, but *how* we use words.

The empirical findings in this paper also resonate with ongoing debates about machine-learned biases around gender. The fact that gendered information is conveyed in context words echoes research on machine-learning bias: that machine-learning models of language, like Google Translate, learn gender-stereotypical meanings of words. Initial solutions to reduce machine-learning gender bias were to blind models to specific gender words (like pronouns). However, recent work shows that this method often just covers up, rather than removes gender biases (Gonen and Goldberg 2019). Gendered meanings are distributed throughout language, rather than in any particular words. This finding resonates with structuralist accounts of language: words are part of a *system* rather than isolated cultural objects.

Chapter 4. Integrating topic modeling and word embedding to characterize violent deaths

ABSTRACT

There is an escalating need for methods to identify latent patterns in text data from many domains. We introduce a method to identify topics in a corpus and represent documents as topic sequences. Discourse atom topic modeling (DATM) draws on advances in theoretical machine learning to integrate topic modeling and word embedding, capitalizing on their distinct capabilities. We first identify a set of vectors ("discourse atoms") that provide a sparse representation of an embedding space. Discourse atoms can be interpreted as latent topics; through a generative model, atoms map onto distributions over words. We can also infer the topic that generated a sequence of words. We illustrate our method with a prominent example of underutilized text: the US National Violent Death Reporting System (NVDRS). The NVDRS summarizes violent death incidents with structured variables and unstructured narratives. We identify 225 latent topics in the narratives (e.g., preparation for death and physical aggression); many of these topics are not captured by existing structured variables. Motivated by known patterns in suicide and homicide by gender and recent research on gender biases in semantic space, we identify the gender bias of our topics (e.g., a topic about pain medication is feminine). We then compare the gender bias of topics to their prevalence in narratives of female versus male victims. Results provide a detailed quantitative picture of reporting about lethal violence and its gendered nature. Our method offers a flexible and broadly applicable approach to model topics in text data.

Digital technologies have produced a deluge of computer-readable text: tweets, blogs, legal documents, product reviews, scientific articles, financial reports, electronic health records, and administrative records (e.g., from public health surveillance). Despite its promise, deriving meaningful information from large-scale text remains a challenge (Hirschberg and Manning 2015). This is especially so in real-world applications, which often put particular demands on methods for computational text analysis. Such methods should be interpretable. They should adapt to the nuances of different discourses. And they should have strong theoretical foundations. In this paper, we offer a new approach that meets these demands: Discourse Atom Topic Modeling (DATM). DATM integrates topic modeling (Blei 2012) and word embedding (Mikolov, Chen, et al. 2013) to identify latent topics in embeddings and map documents onto topics. Methods developed for embeddings (e.g., latent dimensions of cultural meaning Bolukbasi et al. 2016; Garg et al. 2018; Kozlowski et al. 2019) can be applied directly to the topics. We illustrate the value of DATM using text data collected through an ongoing public-health surveillance system for lethal violence in the U.S.

Violent death surveillance provides a striking example of the promise and challenge of computational text analysis. Violent deaths are among the leading causes of mortality in the U.S.(Murphy et al. 2018): More than seven people per hour die a violent death (Centers for Disease Control and Prevention 2019). Understanding and reducing the frequency of these deaths is a major goal for public health. Much of what we know about violent death comes from large administrative databases like the NVDRS, a nationwide public-health surveillance dataset established by the Centers for Disease Control in 2003 (e.g., Ertl et al. 2019; Paulozzi et al. 2004). The NVDRS contains both structured variables (e.g., victim demographics) and unstructured text narratives. These narratives describe the circumstances of death incidents based on reports from law enforcement, medical examiners/coroners, toxicology reports, and crime laboratories. While much has been learned from the NVDRS, researchers have largely used the structured variables; traditional qualitative methods are too labor-intensive to use at scale. The narratives, summarizing more than 300,000 violent deaths, remain mostly unused, despite their potential to illuminate many aspects of violent death, from proximate correlates to nuanced context.

Consider a well-known and durable pattern: differences in violence by gender. Men are more likely than women to die from and perpetrate lethal violence (Batton 2004; Fox and Fridel 2017). Men and women victims also tend to die by different methods (Callanan and Davis 2012; Fox and Fridel 2017). Among suicides and homicides, for example, men are more likely to use firearms, while women are more likely to use poisonous substances (Callanan and Davis 2012; Fox and Fridel 2017). While such gender-linked patterns are reflected across structured variables in the NVDRS (and are well-documented in the literature), the NVDRS *narratives* may also encode gendered patterns — some as yet unidentified. Gendered patterns in text are expected; a growing body of computational work illustrates how and how often information about gender manifests in language (e.g., Bolukbasi et al. 2016; Garg et al. 2018).

The case of violent death surveillance highlights two problems that computational text analysis can solve. First, researchers often want to summarize large corpora, e.g., by extracting major themes like “hot” scientific topics in *PNAS* (Griffiths and Steyvers 2004). Second, researchers want to find evidence for patterns suggested by theory or prior scholarship, e.g., the presence and dynamics of gender and ethnic stereotypes in media language (Garg et al. 2018).

Existing methods can solve both of these problems, but separately. DATM enables us to do both at once. It integrates two major innovations in computational text analysis: topic modeling (Blei 2012) and word embedding (e.g., Mikolov, Chen, et al. 2013).

Topic modeling methods identify latent themes in a corpus and connect those themes to observed words and documents. In conventional topic modeling, topics are distributions over words, and documents are distributions over topics. Powerful as they are, existing topic modeling approaches—especially those commonly used in computational social science—remain largely disconnected from contemporary strategies to represent semantic information using word embeddings. For details and exceptions, see the Supplementary Information (SI).

Word embedding methods represent word meanings by mapping each word in the vocabulary to a point in an *N*-dimensional semantic space (a "word vector"). Words used in similar contexts in the corpus are mapped to nearby points. In a well-trained embedding, word vectors represent semantic information in ways that correspond to human meanings. For example, words that humans rate as similar tend to be closer in semantic space. While word embeddings explicitly model words, they also encode latent semantic structures, like dimensions that correspond to gendered meanings (e.g., Arseniev-Koehler and Foster 2020; Bolukbasi et al. 2016; Garg et al. 2018); analysts can use these dimensions to quantify the latent meanings (e.g., gender) of all the words in a corpus. Topic modeling and word embedding thus have distinct strengths and limitations.

DATM identifies topics (latent themes) and infers the distribution of topics in a specific document, just like a standard topic model. Unlike standard topic modeling, however, DATM does so in an explicit embedding framework; both words and topics live in one semantic space. Our method therefore offers rich representations of topics, words, phrases, and latent semantic dimensions in language. It does so by integrating several theoretical advances to explain word embeddings and efficiently represent sentences in semantic space (Arora et al. 2018, 2017; Arora, Li, Liang, et al. 2016; Ethayarajh 2018).

After describing DATM, we use it to identify key topics in narratives describing over 300,000 violent deaths in the NVDRS (2003-2017). We observe a range of topics, including ones about family, preparation for death, and causality. Using recent approaches to identify semantic dimensions in embedding space (Arseniev-Koehler and Foster 2020; Bolukbasi et al. 2016) we identify a gender dimension and compute the gendered meanings of our *topics*. We describe two illustrative topics in depth: (1) rifles and shotguns and (2) sedative and pain medications. Our approach allows us to summarize and contextualize large-scale, unstructured accounts of violent death. It also allows us to zoom in on "needles" in this haystack of data (Boyd-Graber et al. 2014) and investigate patterns suggested by theory or prior scholarship.

INTEGRATING TOPIC MODELING AND WORD EMBEDDING WITH THE DISCOURSE ATOM TOPIC MODEL (DATM)

To integrate topic modeling and word embedding, we address two core methodological challenges. First, we identify latent topics in a trained word embedding space (also referred to as semantic space); here, we set out to identify topics in an embedding space trained on narratives of violent death. Second, we identify the topic(s) underlying an observed set of words (e.g., a sentence, document, or death narrative). More generally, we need a theoretical framework to connect an embedding space to raw text data. DATM integrates several methodological and theoretical advances in research on word embeddings to address these two challenges, as described next.

Identifying Topics in Semantic Space

We begin with a word embedding trained on a specific corpus (in our case, narratives of violent death). To identify topics in this embedding space, we apply K-SVD, a sparse dictionary learning algorithm (Aharon, Elad, and Bruckstein 2006; Rubenstein, Zibulevsky, and Elad 2008), to the word-vectors (Arora et al. 2018). This algorithm outputs a set of K vectors (called discourse vectors by Arora et al. 2018) such that any of the V word vectors in the vocabulary can be written as a sparse linear combination of atom vectors. Using the generative model below (Eq. 1), atom vectors can be interpreted as topics in the embedding space. The words closest to each atom vector characterize the topic.

We apply K-SVD to our word embedding while varying the number of atom vectors K. To select a final sparse representation, we use a combination of previously proposed metrics for topic model quality and an additional metric suitable for K-SVD (R^2). Together, these metrics quantify 1) how internally coherent topics are; 2) how distinct topics are from each other; and 3) how well the underlying atoms explain the semantic space itself. We select our final model (with 225 topics) to balance performance across these metrics. See SI for details and for a comparison with other topic modeling approaches.

Moving from Semantic Space to Text Data, and Back

Sparse dictionary learning offers a way to identify the "building blocks" of semantic space, but it does not map observed sequences of words (e.g., sentences) to these building blocks. Fortunately, a recently proposed language model offers a link between observed words and points in semantic space: the Latent Variable Model (Arora, Li, Liang, et al. 2016; Arora et al. 2017). This model provides a simplified, probabilistic account for how the text in a corpus was generated. But it also provides a theoretically motivated algorithm to summarize a given set of words as a *context vector* in the semantic space, i.e., a sentence embedding (Arora et al. 2017; Ethayarajh 2018). For a given context vector, we can find the closest atom vector in semantic space, and thus map observed sequences of text data to latent topics. For each document, we assign each context window in a sequence of context windows to a topic. This yields a sequence (or, ignoring order, a distribution) of latent topics that represents the document.

The Latent Variable Model.

Consider first a simplified version of the Latent Variable Model (Eq. 1). The probability of a word w being present at some location *t* in the corpus is based on the similarity between its word vector **w** and the latent "gist" at that point in the corpus $\boldsymbol{c_t}$, i.e., the discourse vector (Arora, Li, Liang, et al. 2016). The word most likely to appear at *t* is the word most similar (closest in semantic space) to the current gist.[25] The similarities between possible word vectors and the

[25] Similarity is measured as the dot product between the two vectors.

discourse vector can be turned into a probability distribution over words by (1) exponentiating the similarities and (2) dividing by their sum Z_{c_t}, so that the distribution sums to 1 (Eq. 1). The gist is latent; $\boldsymbol{c_t}$ is a vector in the semantic space that represents the underlying meaning of the current context. Equation 1 thus associates a distribution over words to every point in semantic space. It also sets up a correspondence between atom vectors (as points in semantic space) and topics. In the generative model (Arora, Li, Liang, et al. 2016), the gist makes a slow random walk through semantic space; at each step *t* a word is emitted according to Equation 1.

$$\Pr[w \text{ emitted at } t|\mathbf{c}_t] = \frac{\exp\left(\langle \mathbf{c}_t, \mathbf{w} \rangle\right)}{Z_{\mathbf{c}_t}}.$$

This simple model is enough to recover many properties of word embeddings (Arora, Li, Liang, et al. 2016).

Arora et. al. (2016) build on Equation 1 to give a more realistic generative model. The conditional probability of a word w being present at some point *t* in the corpus depends on several factors. It depends on the overall frequency of the word in the corpus, p(w). But it also depends on the *local* context or “gist” (the familiar $\boldsymbol{c_t}$), as well as the *global* context of the corpus ($\boldsymbol{c_0}$). The global context vector $\boldsymbol{c_0}$ represents the overall syntactic and semantic structure of the corpus, independent of any local context. The specific combination of local and global context vectors $\boldsymbol{\tilde{c}_t}$ is a linear combination of $\boldsymbol{\tilde{c}_t}$ and $\boldsymbol{c_0}$. The relative importance of word frequency and context is controlled by the hyperparameter α; local and global context trade off with hyperparameter β. This improved Latent Variable Model is written formally in Equation 2 below and detailed elsewhere (Arora et al. 2018, 2017; Arora, Li, Liang, et al. 2016; Ethayarajh 2018). Equation 2 is:

$$\Pr[w \text{ emitted at } t|\mathbf{c}_t] = \alpha p(w) + (1-\alpha)\frac{\exp\left(\langle \tilde{\mathbf{c}}_t, \mathbf{w} \rangle\right)}{Z_{\tilde{\mathbf{c}}_t}},$$

where $\tilde{\mathbf{c}}_t = \beta \mathbf{c}_0 + (1-\beta)\mathbf{c}_t$ and $\mathbf{c}_0 \perp \mathbf{c}_t$.

Mapping observed words into semantic space.

In the generative direction, Equation 2 fixes the probability of a word appearing, given details of the corpus and the current gist. In applications, however, we observe the words; the gist is latent. In DATM, we want to infer the gist (i.e., where we are in semantic space) given an observed set of context words, and then map this gist to an atom vector. Here we summarize work by Arora and colleagues that uses this model to derive a theoretically motivated, high quality embedding of a set of context words: Smooth Inverse Frequency (SIF) embeddings (Arora, Li, Liang, et al. 2016; Arora et al. 2017).

Given the generative model in Equation 2, we can compute the Maximum a Posteriori estimate of the combined context vector $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$ for a set of context words C (see Arora, Li, Liang, et al. 2016; Arora et al. 2017). This is equation 3:

$$(\tilde{\mathbf{c}}_t)_{\mathrm{MAP}} = \sum_{w \in \mathscr{C}} \frac{a}{p(w)+a}\mathbf{w}, \text{ where } a = \frac{1-\alpha}{\alpha Z}.$$

$\tilde{\boldsymbol{c}}_{\boldsymbol{t}}\boldsymbol{MAP}$ is a weighted average of the word vectors in the context window; words are weighted based on their corpus frequency p(*w*). Frequent words make a smaller contribution to the estimate of $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$.[26]

For a given set of context words, we now have an estimate of $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$ (recall that $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$ is a linear combination of local gist and global context for the corpus). But we are fundamentally interested in the local gist $\boldsymbol{c}_{\boldsymbol{t}}$. To recover this, we need an estimate of the *global context* $\boldsymbol{c}_{\boldsymbol{0}}$, which we can then subtract from our estimate of $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$ (see Arora et al. 2017). We first estimate $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$ for a sample of context windows (e.g., sentences) in the data using Equation 3. Then we compute the first principal component of the $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$'s, recovering the direction with the most variance among the context vectors. We interpret this first principal component as the global context vector $\boldsymbol{C}_{\boldsymbol{0}}$.[27] For a given set of context words C, we can estimate $\boldsymbol{c}_{\boldsymbol{t}}$ by using Equation 3 to compute $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}\boldsymbol{MAP}$ from the word-vectors in C, and then subtracting off its projection onto $\boldsymbol{c}_{\boldsymbol{0}}$. The result is an estimate of the latent gist of C, as a point in semantic space.

[26] The amount of re-weighting is controlled by the parameter α. A lower value for α leads to more extensive down-weighting of frequent words, compared to less frequent words. This parameter ranges reasonably from 0.001 to 0.0001; we use a value of 0.001(Arora, Liang, and Ma 2017; Ethayarajh 2018).

[27] After we subtract off the projection of a specific $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$ onto the global context vector $\boldsymbol{c}_{\boldsymbol{0}}$, the remaining vector $\boldsymbol{c}_{\boldsymbol{t}}$ captures the local context---the gist of what is being talked about. We have removed any information in $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$ that corresponds to what the corpus as a whole usually talks about, or how it talks about this (i.e., $\boldsymbol{c}_{\boldsymbol{t}}$ does not just capture frequent words).

Prior work (Arora et al. 2017; Ethayarajh 2018) demonstrates that SIF sentence embedding (i.e., weighting word vectors by frequency, summing them, and removing the global context vector) is also an *empirically* effective representation of the meaning captured by a sentence (or other set of words). In fact, by several metrics, SIF embedding outperforms more sophisticated approaches to represent sentences. Readers familiar with word embeddings may note the correspondence between this MAP and representations of context in the Continuous-Bag-of-Words model (Arora et al. 2017; Mikolov, Chen, et al. 2013); see SI.

SIF embedding allows us to map a set of observed words to a location in semantic space. Given that location, we can find the atom vector that is most similar to this estimated gist, i.e., $\arg\max_{\mathbf{k}\in\mathscr{K}} \cos\left(\mathbf{k}, \mathbf{c}_t\right)$. This atom vector $\mathbf{k}$ then immediately yields the closest *topic* in semantic space.

We have combined three ingredients — sparse coding of the embedding space (Arora et al. 2018), the Latent Variable Model (Arora, Li, Liang, et al. 2016), and sentence embeddings (Arora et al. 2017) — into a cohesive procedure that allows researchers to discover latent topics in a corpus and to identify the topic that best matches the estimated gist of an observed context window. Finally, to infer topics across an entire document, we estimate the gist $\boldsymbol{c_t}$ over rolling context windows in the document.[28] This is consistent with a key assumption of the Latent Variable Model: That the gist changes slowly across a document. This last step yields the sequence (or, ignoring order and dividing the topic counts by a normalizing constant, the

[28] Here, we use a context window size with 10 terms

distribution) of topics underlying the document.[29] Here, we code topics as binary variables for each record (present or not).[30] Taken individually, each component of DATM offers an effective tool for specific tasks and analyses. Once integrated, they generate a strikingly effective and general approach to analyze real-world text data.

TOPICS IN DESCRIPTIONS OF VIOLENT DEATH

Our data are drawn from the NVDRS from 2003 to 2017 (Paulozzi et al. 2004). This NVDRS database included information about 307,249 violent deaths forwarded from 34 U.S. states and the District of Columbia. This state-level information is abstracted into the NVDRS by public health workers (PHW) using a standardized codebook. We use two text variables in the NVDRS

[29] Turning the sequence into a distribution accounts for word count differences; it also provides a connection to traditional topic modeling, which associates a distribution over topics to each document.

[30] Note that our approach to assigning topics is fundamentally different from the approach in traditional topic modeling; rather than trying to model each document as a mixture of topics, DATM models each document as a trajectory in semantic space, and (in parallel) decomposes semantic space into topical building blocks. The topic closest to each point of the trajectory is then assigned to the document. For this reason, DATM does not take a document-term matrix as input. In a sense, DATM provides a more “bottom-up” approach to inferring topics, rather than a “top down” approach involving further assumptions about the role of topics in the generative process.

written by PHW: narratives of 1) law enforcement reports and 2) medical examiner or coroner investigative reports. Death records may include one of these variables, both, or none, for a total of 568,262 narratives. We train our word embedding on all of these narratives using word2vec (Mikolov, Chen, et al. 2013). After applying several exclusion criteria, our final sample is 272,964 deaths. For details, see the SI and our code: https://github.com/arsena-k/discourse_atoms. For data access, apply to The Centers for Disease Control and Prevention: https://www.cdc.gov/violenceprevention/ datasources/nvdrs/dataaccess.html.

When we applied DATM to the NVDRS narratives, the resulting 225 topics covered various aspects of violent death. For example, we observed several topics about weapons, substance use, and forensic analyses. To interpret a given topic, we examine the 25 terms closest to the topic's atom vector and then we assign the topic a label using face validity. We list several topics in Table 3 and all topics in the SI.

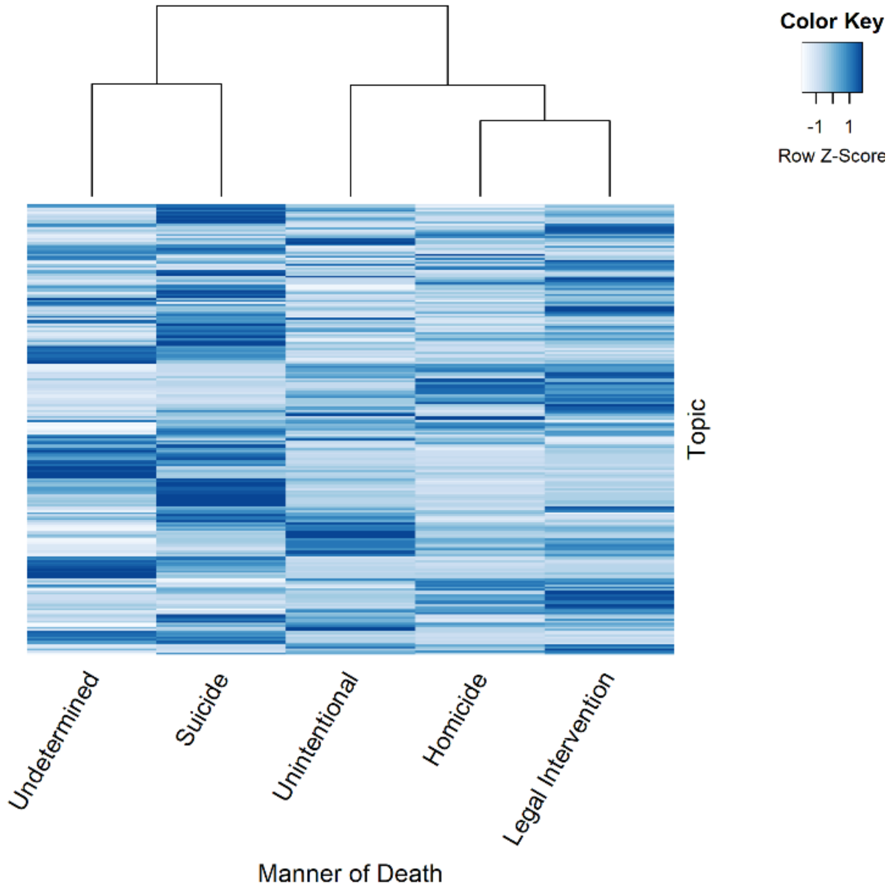

Figure 5. Prevalence of 225 Topics in Narratives of 272,964 Decedents of Violent Death, by Manner of Death. Note: Each row represents the fraction of narratives with a given topic by manner of death, row standardized across all manners of death.

Figure 5 illustrates the prevalence of our 225 topics as patterned by manner of death: suicide, unintentional shooting, homicide, homicide resulting from legal interventions (e.g., police shootings), and deaths of undetermined intent. Each row represents the fraction of narratives with a given topic, by manner of death. The dendrogram[31] shows that across the manners of death, suicides are most similar in topic distributions to undetermined deaths; this

[31] Computed using hierarchical clustering with Euclidean distance for a similarity metric.

makes sense, because many deaths may look like suicide but lack the required evidence for classification as suicide (Rosenberg et al. 1988). It also shows that homicides are most similar to legal intervention deaths, reflecting that legal intervention deaths are a unique *type* of homicide.

Table 3. Sample Topics within Narratives of Violent Death.

Topic Label	**Seven Most Representative Terms**
Physical Aggression	tackled, lunged_toward, began_attacking, advanced_toward, attacked, slapped, intervened
Causal Language	sparked, preceded, triggered, precipitated, led, prompted, culminated
Preparation for Death	disposal, deeds, prepaid_funeral, burial, worldly, miscellaneous, pawning
Cleanliness	unkempt, messy, disorganized, cluttered, dirty, tidy, filthy
Everything Seemed Fine	fell_asleep, everything_seemed_fine, seemed_fine, wakes_up, ran_errands, ate_breakfast, watched_television
Suspicion and Paranoia	conspiring_against, plotting_against, restraining_order_filed_against, belittled, please_forgive, making_fun, reminded
Reclusive Behavior and Chronic Illness	recluse, heavy_drinker, very_ill, chronic_alcoholic, bedridden, reclusive, recovering_alcoholic

Notes: Most representative terms are listed in order of highest to lowest cosine similarity to each topic's atom vector. Topic labels are manually assigned. As part of preprocessing the narratives, we transformed commonly occurring phrases into single terms (Řehůřek and Sojka 2010).

Topics and Latent Semantic Dimensions

Because the atom vectors corresponding to topics live in an embedding space, we can apply common word embedding methods to our topics. One prominent deductive approach uses knowledge about cultural connotations to extract a corresponding dimension in the semantic space. Here, we extract a dimension for gender (masculine vs feminine) in the corpus, following standard word embedding methods (e.g., Bolukbasi et al. 2016).[32] We then examine the topics that load most highly onto the gender dimension (i.e., have the highest or lowest cosine similarity). Cosine similarity can range from -1 to 1: for gender, the topics with large negative cosine similarity are more distinct to language about men (and not women), while topics with large positive cosine similarities are more distinct to language about women (and not men).

In our data, the most masculine topic is about the military, followed by topics about rural outdoor areas, rifles and shotguns, specific outdoor locations, and characteristics of suspects. The most feminine topic, by contrast, is one about sedative and pain medications, followed by topics about poisoning, children, drug concentrations, and psychiatric medications. Surprisingly, we also observe that a topic about games is highly gendered (i.e., the seventh most masculine topic). This topic reflects a range of games, including video or computer games. Prior work highlights games or forms of play linked to violent death (e.g., russian roulette, choking games, children playing with guns (Hemenway, Barber, and Miller 2010)); the fact that this topic is highly masculine suggests that such deaths may be distinctly patterned by gender.

In Figure 6, we compare the *cosine similarity* of topics to this gender dimension with the mean *prevalence* of each topic among female victims (versus among male victims). These two variables capture distinct ways that gender is encoded in the NVDRS, which we might expect to be strongly related. Similarity to the gender dimension reflects the appearance of topics in the

[32] To extract a gender dimension, we average the vectors for the words: woman, women, female, females, she, her, herself, and hers. We then subtract out the average of the vectors for the words: man, men, male, males, he, him, himself, and his (Arseniev-Koehler and Foster 2020; Bolukbasi et al. 2016).

context of gendered *language* in the narratives; it can reveal gendered patterns in topics even when there is no corresponding metadata for documents. Mean prevalence captures the extent to which a topic is mentioned among men versus women victims. The strong correlation (Spearman $\rho = 0.69$, $p<0.0001$) suggests that topics are gendered in semantic space in a way that indeed corresponds to the gender of the victims in the narratives.

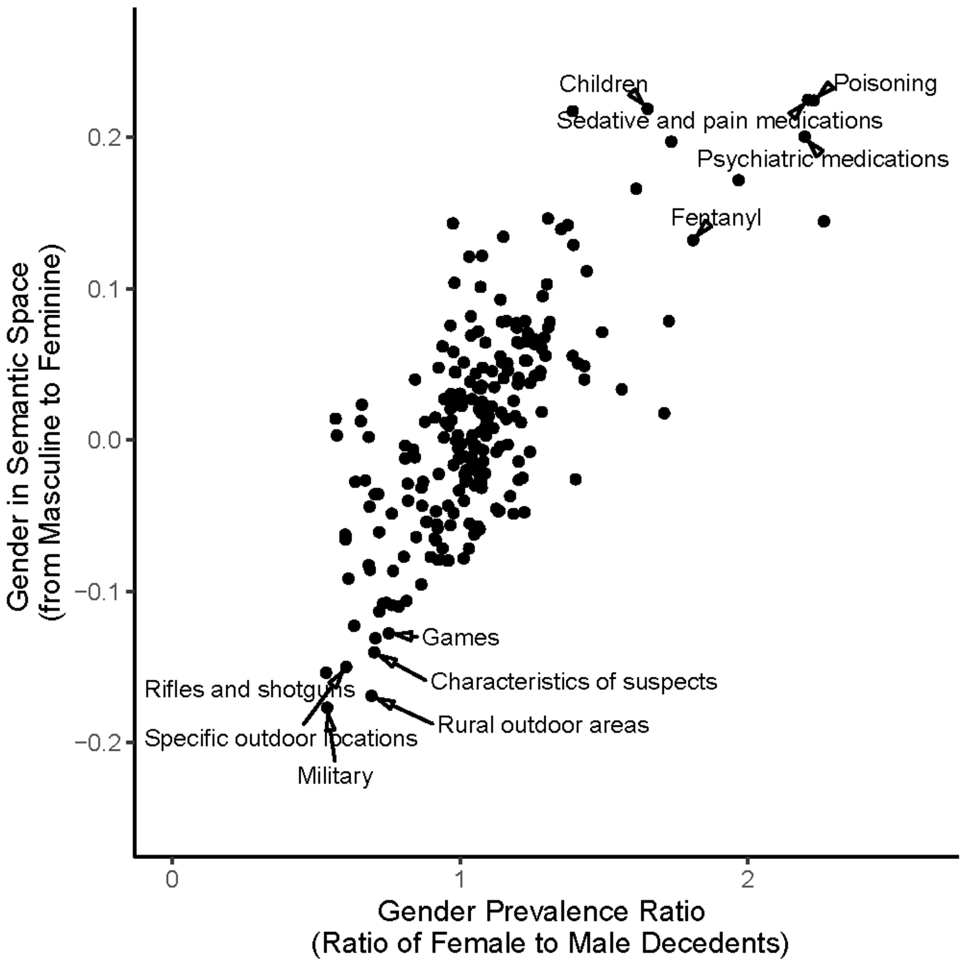

Figure 6. Latent Gendered Meanings of Topics vs Prevalence of Topics in Female vs Male Decedents' Narratives. Notes: N= 225 topics. For clarity, labels are shown only for topics with high or low gender meanings or gender prevalence ratios; overlapping labels are removed. The y axis represents cosine similarity between a given topic and the gender dimension in semantic space. The x axis represents the ratio of female decedents' narratives containing a given topic compared to narratives of male decedents.

Next, we describe two topics in depth. Each has a high cosine similarity to the gender dimension. We select these topics because they are the most masculine and feminine topics (respectively) that describe weapons of death; weapon-use has a well-known gendered pattern in violent death (e.g., Callanan and Davis 2012; Fox and Fridel 2017). For each topic, we describe the most representative words and the case that loads most highly onto it. Then we use logistic regression to describe correlates of the topic: decedent demographics, manner of death, and number of decedents in the incident, controlling for word count.

Topic 141: rifles and shotguns.

Topic 141 reflects characteristics of long guns (e.g., rifles and shotguns). These firearms are typically owned for hunting and sport shooting (Hepburn et al. 2007) and can be used to shoot at far ranges (compared to handguns). The most representative terms refer to makes and models of long guns, as well as characteristics of gun action: how the gun is loaded and fired. The highest loading case describes the death of a young man accidentally shot by a friend playing with a rifle, who believed it was unloaded. The narrative describes the gun in depth (e.g., as a “bolt action deer rifle”).

Topic 141 is the third most masculine topic in semantic space. This strong gender connotation reflects the fact that violent death by firearm typically involves males (Callanan and Davis 2012; Fox and Fridel 2017). Logistic regression confirms that Topic 141 is distinctly more common among male victims (than females), controlling for characteristics listed in Table 4 (adjusted odds ratio= 0.49, 95% CI: 0.48-0.51). The strong gendered associations of this *particular* gun-related topic in semantic space (compared to say, Topic 61: Handguns) could follow from the fact that far more men than women own long guns (Hepburn et al. 2007).

We also observe patterns in Topic 141 across other covariates. While prior work suggests that the majority of firearm-related decedents are Black (Goldstick, Carter, and Cunningham 2021), our results in Table 4 suggest that patterns may be more nuanced for deaths involving long guns. For instance, this topic is more common among American Indian/Alaska Native decedents, and less common in all other race/ethnicity groups (compared to Whites). Finally, the topic was more common in incidents where there were multiple deaths, as one would see in mass shootings. Findings from this topic underscore the need for work on specific guns (e.g., Hanlon et al. 2019) in order to more effectively target prevention efforts aimed at firearm control.

Topic 53: sedative and pain medications.

Topic 53 involves sedatives and medications that can be used to control pain. The most representative terms for this topic refer to the names of such medications (e.g., "phenergan"). The highest loading case describes a middle-aged white male decedent who was found dead next to various prescription bottles with pain medications (e.g., methadone and hydrocodone). The immediate cause of death was ruled as suicide. In general, we found many topics focused on distinct groups of medications and drugs, attesting to the depth and patterned ways in which substances are described in the narratives.

Topic 53 is the most feminine topic in semantic space. This strong feminine connotation may reflect the fact that women are more likely to die by poisoning in suicide (Ertl et al. 2019). Logistic regression confirms that Topic 53 is distinctly more common among female victims, controlling for characteristics listed in Table 4 (adjusted odds ratio= 2.52, 95% CI: 2.47-2.58). We observe additional patterns of topic prevalence across these correlates. Compared to suicides, this topic is more common in undetermined deaths, but less common in all other deaths. The fact that unclassified deaths disproportionately involve this topic in their narratives resonates with broader scholarship on the misclassification of manner of death. Undetermined deaths are predominately associated with drug intoxication and poisoning (Rockett et al. 2018), and many undetermined deaths involving drugs may be uncounted suicides (e.g., Stone et al. 2017).

Table 4. Characteristics of Violent Deaths with Two Selected Topics

	Topic	
	Rifles and Shotguns	**Sedative and Pain Medications**
Characteristic	AOR (95% CI)	AOR (95% CI)
Female Decedent[1]	0.49 (0.48-0.51)	2.52 (2.47-2.58)
Decedent Race/Ethnicity[2]		
American Indian/Alaska Native, NH	1.31 (1.20-1.42)	0.46 (0.41- 0.52)
Asian/Pacific Islander, NH	0.48 (0.43-0.54)	0.64 (0.59-0.70)
Black or African American, NH	0.88 (0.85-0.91)	0.54 (0.51- 0.56)
Hispanic	0.59 (0.56-0.62)	0.63 (0.60-0.67)
Two or more races, NH	1.01 (0.92-1.10)	0.80 (0.73-0.88)
Unknown race, NH	0.70 (0.56-0.87)	0.70 (0.56-0.87)
Decedent Age (Years)[3]		
20-29	0.96 (0.91-1.00)	1.37 (1.29-1.46)
30-39	0.90 (0.86-0.95)	1.74 (1.64-1.85)
40-49	0.93 (0.88-0.98)	1.97 (1.86-2.10)
50-59	1.03 (0.98-1.08)	2.17 (2.04-2.30
60+	1.40 (1.33-1.47)	1.68 (1.58-1.79)
Manner of Death[4]		
Homicide	0.79 (0.77-0.82)	0.14 (0.13-0.15)
Legal Intervention	1.09 (1.01-1.17)	0.22 (0.19-0.26)
Undetermined	0.06 (0.06-0.07)	2.01 (1.95-2.07)
Unintentional	3.16 (2.84-3.51)	0.13 (0.10-0.19)
Multiple Decedents in Incident[5]	1.76 (1.68-1.84)	0.40 (0.37-0.43)
Word Count[6]	1.00 (1.00-1.00)	1.00 (1.00-1.00)
Notes: N = 272,964 decedents. Topics are coded as present in any amount or not (1/0) in either the narrative of law enforcement reports or of medical examiner/coroner reports. AOR = Adjusted Odds Ratio; CI = Confidence Interval; NH=Non-Hispanic. [1]Referent = Male. [2]Referent = non-Hispanic White. [3]Referent = 12-19. [4]Referent = Suicide. [5]Referent= Incidents with a single decedent. [6]This is the combined word count of both narratives.		

These results illustrate that the same methods used to identify the biases or cultural meanings of *words* in word embeddings can also be used to identify biases of *topics* extracted with DATM. These methods extend to semantic dimensions beyond gender (e.g., Arseniev-Koehler and Foster 2020; Bolukbasi et al. 2016); we provide another example (outdoors versus indoors) in the SI.

DISCUSSION

In this paper, we introduced a new method to model topics: Discourse Atom Topic Modeling (DATM). In DATM, topics, sentences, words, and other semantic structures are all represented in a single semantic space. Raw text can be mapped into this space to distill individual documents into sequences of topics and thus draw out the prevalence of topics in a corpus. Using DATM, we discovered a range of themes buried in descriptions of lethal violence from a large administrative health dataset. We observed that the gendering of these topics in semantic space corresponds to the ratio of female versus male victims whose narratives mention these topic, and analyzed two highly gendered topics in depth. Methodologically, our model builds on theoretical work to explain word embeddings and represent sentences in embedding spaces (Arora et al. 2018, 2017; Arora, Li, Liang, et al. 2016), as well as a wealth of previous models to extract topics (Blei 2012; Griffiths and Steyvers 2004).

For computational social science and natural language processing, DATM provides a new, integrated approach to discover patterns in large-scale text data. As a topic model, DATM picks up fine-grained, interpretable topical structures. These topics are coherent despite the fact that no stopwords were pre-specified. This makes DATM ideal for real-world applications of text analysis, which are often domain specific and would otherwise require specialized lists of stopwords. Further, DATM offers a cohesive, theoretically-motivated approach to *integrate* questions that are often asked with topic models with questions often asked with embedding methods. A researcher can now ask, for example, not only what topics are in a corpus, but how these topics lie on latent semantic dimensions such as gender or social class.

For public health, our results illustrate patterns encoded in large-scale narrative data about suicides, homicides, and other violent deaths. Using DATM on these data offers a new way to break out of the well-worn categorical systems by which we interpret and monitor lethal violence.

We found that unstructured text data can hide potential patterns or trends that are not yet part of our standardized menu of structured variables. Such patterns could suggest new lines of research that aim to reduce violent death; for example, discovering new indicators of suicide risk, with eventual implications for medical providers or hotlines. Despite the wide use of the NVDRS for research and policy about lethal violence, actionable information in its text data has remained largely out of reach. We hope that DATM will provide an interpretable, flexible, theoretically grounded, and effective tool for scientists to unlock the potential of important datasets like the NVDRS.

SUPPLEMENTARY INFORMATION

Data

As described in the main text, our data are drawn from the National Violent Death Reporting System (NVDRS) collected between 2003-2017. Here we provide additional details on these data. The Centers for Disease Control and Prevention (CDC) share these restricted data with researchers via the execution of a standard use agreement. Users may apply to the CDC directly for data access: https://www.cdc.gov/violenceprevention/ datasources/nvdrs/dataaccess.html.

These data include 307,249 violent deaths, for decedents aged 12 and older: suicides (N=192,115), homicides (N=73,602), deaths of undetermined intent (N=34,266), and 7,266 other deaths such as unintentional deaths (primarily, shootings) and deaths related to legal intervention (e.g., police shootings). Legal intervention deaths are defined using criteria outlined by Barber et. al. (2016). As described in the main text, each death record may be accompanied by a narrative of medical examiner/coroner reports, a narrative of law enforcement reports, both narratives, or neither. In total, these data include 302,072 narratives of medical examiner/coroner reports, and 266,190 narratives of law enforcement reports. Initial cleaning of the narratives included corrections for misspelling and minor editing for common abbreviations (e.g., COD: “cause of death”). In the case of multiple death incidents, when the narrative referred to the current victim, “victim” was recoded as “primary_victim” and all other mentions of “victim” were recoded as “extra_victim.” We also transformed commonly occurring phrases into single words (i.e., terms) based on collocation (Řehůřek and Sojka 2010). The medical examiner/coroner narratives had an average length of 105 terms (SD=77); the law enforcement narratives averaged 120 terms (SD=117). Our resulting corpus from these text variables and pre-processing steps included a vocabulary size of 28,222 unique terms. During training of our embedding (described in SI), we removed terms that are in the dataset fewer than 15 times to avoid learning low quality word-vectors for these terms. As described in the main text, we coded topics as binary variables for each death record (present in any amount in either the medical examiner/coroner narrative or the law enforcement narrative=1, not present in any amount=0).

We limit our empirical investigation of topic distributions to the 272,979 (88.85%) deaths which have at least 50 terms in either of the narratives. We further exclude 18 deaths where manner of death is missing or is coded as terrorism. This process leaves 272,964 deaths.

In our empirical analyses of the distribution of topics, we used several structured variables in the NVDRS: victim sex (male/female), age at time of death (in years), race/ethnicity, manner of death (suicide, homicide, legal intervention death, undetermined, or unintentional death), and number of victims (1 vs. more than 1). We also used word count of the narrative(s); for cases with narratives of medical examiner/coroner reports as well as narratives of law enforcement reports, word count is combined across both narratives. We coded age into six groups: 12-19, 20-29, 30-39, 40-49, 50-59, and 60 and older. There was no missing data for age or for number of victims. We coded race/ethnicity as: American Indian/Alaska Native, non-Hispanic; Black or African American, non-Hispanic; Hispanic; Two or more races, non-Hispanic; White, non-Hispanic; Asian/Pacific Islander, non-Hispanic; and Unknown race, Non-Hispanic. There was no other missing data in race/ethnicity. To account for missing data for victim sex (N=2), we manually imputed victim sex using information about victim sex described in the narratives (e.g., “victim was a 20 year old male.”). See Table 5 for descriptive summaries of these variables.

Table 5. Characteristics of Sample of Violent Deaths, drawn from the National Violent Death Reporting System.

Characteristic	**N (%) or Mean (SD)**
Female Decedent[1]	64,404 (23.59%)
Decedent Race/Ethnicity	
White, NH	190,474 (69.78%)
American Indian/Alaska Native, NH	4,044 (1.48%)
Asian/Pacific Islander, NH	4,616 (1.69%)
Black or African American, NH	49,218 (18.03%)
Hispanic	19,627 (7.19%)
Two or more races, NH	4,129 (1.51%)
Unknown race, NH	857 (0.032%)
Decedent Age (Years)	
12-19	18,029 (6.60%)
20-29	61,767 (22.63%)
30-39	49,113 (17.99%)
40-49	52,431 (19.21%)
50-59	46,903 (17.18%)
60+	44,721 (16.38%)
Manner of Death	
Suicide	173,006 (63.34%)
Homicide	62,751 (22.99%)
Legal Intervention	5,124 (1.88%)
Undetermined	30,598 (11.21%)
Unintentional	1,485 (0.54%)
Multiple Decedents in Incident[2]	13,991 (5.12%)
Word Count[3]	226.37 (158.72)

Notes: N = 272,964 decedents. SD= Standard Deviation, NH=Non-Hispanic. [1]Referent =Male. [2]Referent= Incidents with a single decedent. [3]This is the combined word count of the narrative of law enforcement reports and the narrative of medical examiner or coroner investigative report.

Training the Word Embedding

We trained our word embedding using word2vec with Continuous-Bag-of-Words (CBOW) and negative sampling. We did so because of the connections of this architecture to our topic modeling approach (see Arora, Li, Liang, et al. 2016; Arora et al. 2017); however, any embedding algorithm can be used to train a semantic space given text data. Using Gensim

(Řehůřek and Sojka 2010) in Python to train our word embedding, we tuned two hyperparameters: dimensionality of the semantic space and context window size. Specifically, we trained word embeddings with 50, 100, 200, and 300 dimensions, training three models at each dimensionality to vary context window size between 5, 7 and 10, for a total of 12 embeddings. We report context window size as the number of words on each side of the target word. Thus, a context window size of 5 means that we use 5 words to the left and 5 words to the right of the target word, for a total of 10 words in the context window C; in general, a context window size *n* implies a context window C with *2n* words total. These hyperparameters are within the range of standard choices for hyperparameters (Pennington et al. 2014).

We selected our final word embedding (200-dimensions and a context window size of 5) by comparing the performance of the 12 different embeddings on two common metrics for assessing the quality of embedding models: the WordSim-353 Test and the family section of the Google Analogy Test. The WordSim-353 Test (Finkelstein et al. 2002) compares the cosine similarity of two words in a word embedding model with similarity assigned by human annotators; our trained word embedding yielded a Spearman correlation of 0.45 (p<0.0001) with human-rated similarities. The Google Analogy Test (Mikolov, Chen, et al. 2013) tests how well an embedding model can complete a series of analogies, divided up in various sections (e.g., family, currency, tense, world capitals). We focused on the family section which is most relevant to our data domain. Our trained word embedding correctly completed 70% of the analogies in the family section. We observe that performance on these metrics varied little across our hyperparameters. We set the number of iterations at 10, and negative samples at 5, and we randomly shuffle the order of the documents prior to training our embedding to prevent any ordering effects.

Connections between Continuous-Bag-of-Words (CBOW) Word Embeddings and DATM

Word2vec learns a semantic space from a corpus by giving a task to an artificial neural network. In Word2vec with CBOW, the task is to guess words from their contexts in the data (i.e., short excerpts of text data, also called context windows). More precisely, for each context window in the data, a CBOW network (CBOW for short) is asked to predict the most likely word (i.e., target word), given the average of the words in the context window (i.e., the context vector).

This is done across the many possible context windows of data, until CBOW reaches a certain level of accuracy in predicting words. Below, let w_t be the target word (the word at "time", or, equivalently, text position, t), with vector $\mathbf{w}_\mathrm{t}$, and let $c_t = \{w_{t-n}, w_{t-n+1}, \ldots, w_{t-1}, w_{t+1}, w_{t+2}, \ldots, w_{t+n}\}$ be a set of words within a context window *C* with size n (i.e., 2n words total).[33] Note that we use bold to distinguish vectors from their corresponding entities; so the word-vector $\mathbf{w}_\mathrm{t}$ corresponds to the word w_t. Given a set of context words c_t, the probability that CBOW will predict word w_t is given by Equation 4:

$$P(w_t|c_t) \propto \exp(\langle \mathbf{w}_t, \bar{\mathbf{c}}_t \rangle), \text{ where } \bar{\mathbf{c}}_t = \frac{1}{2n} \sum_{i=t-n, i \neq t}^{t+n} \mathbf{w}_i.$$

[33] Note that the implementation of CBOW draws the vector for the context words and the vector for the target word from two different weight matrices. The first vector comes from (averaging) the weights linking the input in CBOW's artificial neural network to the hidden layer. The second vector comes from the weights linking the hidden layer to the output layer. See Rong (2014) for a more detailed explanation of the implementation and hyperparameters.

CBOW training adjusts the weights so as to maximize the probability of the actual word corresponding to a given context window, for all word/context pairs. In practice, however, CBOW is trained with two tricks: negative sampling and sub sampling (Mikolov, Yih, and Zweig 2013). These tricks effectively down-sample more frequently occurring context words (Arora et al. 2017). Such techniques *implicitly re-weight* the context words, such that a word is guessed from a weighted sum of its context words, where these weights are based on word frequency. This means that in most practical implementations of CBOW (including the one we use), the “context” vector is computed in the same way as context or “gist” in the Discourse Atom Topic Model (c_t), including the down-weighting of frequent context words. Put another way: practical implementations of CBOW learn a semantic space by predicting the most likely word from the estimated “gist” (Arora et al. 2017), a weighted linear combination of the word vectors. More broadly, this connection implies that CBOW with negative sampling and the Discourse Atom Topic Model form a single cohesive theoretical model. In practice, however, any word embedding can serve as input to the Discourse Atom Topic Model.

Identifying DATM Topics in a Word Embedding with K-SVD

Here, we describe how the K-SVD algorithm works to identify topics in a trained embedding. As described in the main text, this algorithm outputs a set of *K* vectors (called “discourse atoms” by Arora et. al. (2018)) such that any of the *V* word vectors in the vocabulary can be written as a sparse linear combination of these vectors. We refer to these vectors as “atom vectors.” As described next, these atom vectors can be interpreted as topics in the embedding space. The words closest to each atom vector typify the topic. K-SVD is a well-established method (Aharon et al. 2006) and we implement K-SVD using the ksvd package in Python (Rubenstein et al. 2008). Here, we provide details on the K-SVD algorithm to keep the exposition self-contained.[34]

The input to K-SVD is a matrix **Y** of *V* word-vectors, each of which is *N*-dimensional. Thus, this matrix has *N* rows and *V* columns. The goal of applying K-SVD to this matrix is to represent each word-vector as a sparse linear combination of atom vectors, where there are a total of K possible atoms and each is represented by an *N*-dimensional vector. K-SVD output includes two components.

First, the output provides a matrix **D** of atom vectors (which we will ultimately interpret as topics), commonly called the dictionary. **D** has *N* rows and *K* columns; each column is an *N*-dimensional vector corresponding to an atom in the embedding space. Because these atoms are simply vectors in the same semantic space as word vectors, we can compare them to other vectors (like word-vectors, or latent semantic dimensions) in this space using cosine similarity. To understand what a given atom vector represents, we look at the words in the vocabulary whose word vectors have the highest cosine similarity to each atom vector. Note that, under the Latent Variable Model (described in the main text), these word vectors also give the words most likely to be "emitted" when the context coincides with the atom; this allows us to turn each atom vector into a full-blown topic (i.e., probability distribution over words) as in conventional topic modeling.

[34] Throughout we follow the notation and approach of the excellent Wikipedia exposition as well as the original paper (Aharon, Elad, and Bruckstein 2006), fleshing details out and specializing the exposition to our specific case; see: https://en.wikipedia.org/wiki/K-SVD.

Second, the algorithm produces a sparse matrix of coefficients **X** with *K* rows and *V* columns. Each column in this sparse matrix indicates how a given word can be reconstructed as a linear combination of atom vectors: which atom vectors to use and in what amounts. While we do not use them in this paper, these coefficients could be used to see which words load onto a given topic and with what strengths. Arora et al (2018) use these coefficients to disentangle the multiple meanings of words.

Note a key difference between DATM and the more familiar LDA topic modeling: LDA topic modeling decomposes a document-term matrix to find topics; DATM decomposes the embedding matrix. Thus, our approach identifies topics *in the semantic space* of a corpus.

If the output of K-SVD is a good solution, then each word-vector should be well-approximated as a sparse linear combination of atom vectors (i.e., one with few non-zero coefficients). Put another way, using our topics, we should be able to roughly reconstruct the original meanings of the word-vectors. To reconstruct our matrix of word-vectors, we multiply the atom matrix (**D**) by the coefficient matrix (**X**). To find a good representation of the original word vectors, we want to minimize the difference between **Y** (our word vectors) and **DX** (our sparse reconstruction).

Comparing the reconstructed matrix **DX** to the original embedding matrix **Y** yields measures of error in a discourse atom solution (e.g., sum of squared errors, root mean square error, and even R^2). The approximate decomposition is visualized in Figure 7.

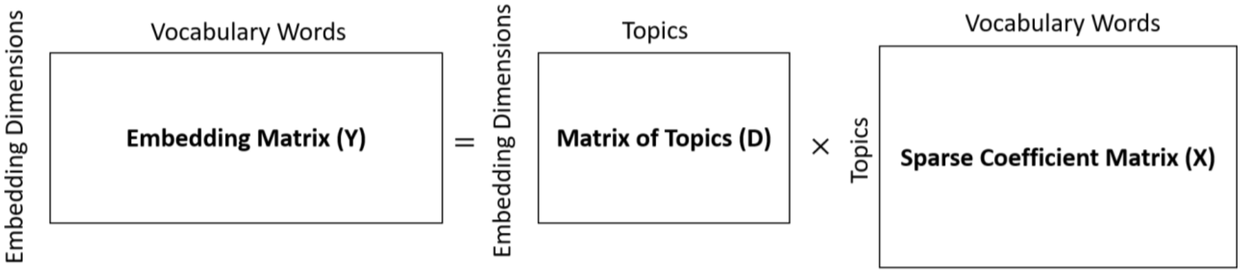

Figure 7. Decomposing the Embedding Matrix into a Dictionary of Topics and Coefficients.

At the same time, we want a sparse solution; that is, we want to make sure that each word is represented by a small number of topics. Formally, we want to keep the ℓ^0 "norm" of each column in **X** (i.e., the number of non-zero elements) small, so that it is less than or equal to the sparsity constraint hyperparameter. Thus, the objective function of the K-SVD constrained optimization problem is:

$$\min_{\mathbf{D},\mathbf{X}} \left\{ \|\mathbf{Y} - \mathbf{DX}\|_F^2 \right\}$$

with the constraint $\|x_i\|_0 \leq T_0$ ∀i. Recall that $\|x_i\|_0$ is the $\ell 0$ norm of the ith column of X and $\| \ldots \|_F^2$denotes the Frobenius norm, i.e., the sum of squared entries of the matrix. Hence we want to choose **D** and **X** such that the total squared difference between the original embedding **Y** and the reconstruction **DX** is minimized, while constraining each column of **X** to T_0 non-zero entries; in other words, a sparse representation of each word vector in terms of the atom vectors.

Solving the objective function of K-SVD to arrive at topics.

In general, this constrained optimization problem cannot be "solved" (i.e., truly optimized); therefore approximate methods must be used. The overall strategy to minimize the objective function of K-SVD (and thus identify our topics) involves alternating updates to the coefficient matrix **X** and the dictionary **D**. We begin with a randomly initialized dictionary **D**.

Updating the coefficients.

Given a fixed dictionary, finding the coefficients is basically a least squares problem: we need to find a distinct, sparse linear combination of atom vectors that best represents each word-vector (i.e., each column of the embedding matrix). In K-SVD, this problem is commonly solved (heuristically) with orthogonal matching pursuit (OMP): a greedy algorithm that iteratively finds a sparse representation for each word vector, where the number of atom vectors allowed is determined by T_0 (Aharon et al. 2006; Pati, Rezaiifar, and Krishnaprasad 1993). The use of OMP exploits the fact that the minimand $\|\boldsymbol{Y} - \boldsymbol{DX}\|_F^2$ can be rewritten as $\sum_i^N \|\boldsymbol{y_i} - \boldsymbol{Dx_i}\|_2^2$ (note the shift from Frobenius to the familiar ℓ^2 norm). Each of the terms in this sum can be separately minimized with respect to the coefficients $\boldsymbol{x_i}$ that correspond to the reconstruction of word vector $\boldsymbol{y_i}$ (with the familiar sparsity constraint T_0 on the number of non-zero coefficients). These separate minimization problems can be addressed using OMP to give an approximate solution (Aharon et al. 2006).[35]

[35] See https://en.wikipedia.org/wiki/Matching_pursuit for a simple exposition of the related Matching Pursuit algorithm. OMP works in our case as follows: For a given word vector, we find the closest possible atom vector using cosine similarity. The projection of the word vector onto that first atom vector represents our first attempt at reconstructing the word vector, and hence our first pass at the coefficients. We next compute the residual (the vector difference between the word vector and the reconstruction). We then find the atom vector closest to the residual (i.e., what is not explained by the atom(s) already assigned to this word-vector). This becomes the next atom vector with a non-zero coefficient. In OMP, we compute new coefficients for both

Updating the dictionary.

Once the coefficients are updated for all columns of **X'**, we freeze the coefficients. We then update the dictionary of atoms; here we follow Ahron et. al. (2006) closely. We update one atom vector (i.e., column of the dictionary) at a time. To update the *k*th atom vector, we identify the word vectors whose reconstructions use that atom (i.e., the corresponding coefficient in the sparse representation vector $\boldsymbol{x_i}$ is nonzero). Now define a representation error matrix $\boldsymbol{E}_k = \boldsymbol{Y} - \sum_{j \neq k} \boldsymbol{d}_j x_T^j$, where $\boldsymbol{d_j}$ is the jth column of the dictionary matrix **D** (i.e., the atom vector for topic *j*) and $\boldsymbol{x}_T^j$ is the *j*th *row* of the representation matrix **X**, i.e., all of the coefficients for the *j*th atom vector. $\boldsymbol{E_k}$ essentially corresponds to all of the reconstruction error that remains after we have reconstructed **Y** with the other *K*-1 topics.

We want to reduce the reconstruction error further by updating the vector for the kth atom $\boldsymbol{d_k}$ and the corresponding row of the coefficient matrix $\boldsymbol{x}_T^k$, but we must do so in a way that preserves sparsity. We do so by considering only the columns of the error matrix that correspond

atom vectors by projecting the full word-vector onto their span (in this case, a plane); this yields a new set of coefficients and a better reconstruction of the original word-vector. We iterate this process---compute the difference between the word-vector and its current reconstruction; find the atom vector closest to the residual; project the full word-vector onto the span of the iteratively selected atom vectors; repeat---until we have chosen T_0 atom vectors, corresponding to T_0 non-zero coefficients in $\boldsymbol{x'}_{\boldsymbol{i}}$ for the sparse coefficient matrix **X'** corresponding to the current dictionary **D'**.

to word vectors whose reconstruction currently uses the kth atom, yielding a restricted matrix $\boldsymbol{E}_k^R$. We likewise restrict $\boldsymbol{x}_T^k$ to only those elements of the row with non-zero entries (i.e., those coefficients where atom vector k is currently used); call this $\boldsymbol{x}_R^k$. We now update $\boldsymbol{d_k}$ and $\boldsymbol{x}_R^k$ to minimize $\left\|\boldsymbol{E}_k^R - \boldsymbol{d_k}\boldsymbol{x}_R^k\right\|_F^2$; this is, in essence, the "best we can do" to further reduce error by only changing the atom vector $\boldsymbol{d_k}$ and altering the way that reconstructions *already using* that atom vector load onto it. By construction, this update cannot lead to violation of the sparsity constraint. This sparsity-preserving minimization with respect to $\boldsymbol{d_k}$ and $\boldsymbol{x}_R^k$ can be done via singular value decomposition (SVD) of the error matrix $\boldsymbol{E}_k^R = \mathbf{U}\,\boldsymbol{\Delta}\,\boldsymbol{V}^T$. In essence we want a rank one approximation of the error matrix $\boldsymbol{E}_k^R$; the optimal such approximation is obtained by setting $\boldsymbol{d_k}$ to be the first left singular vector (the first column of **U**) and the reduced coefficient vector $\boldsymbol{x}_R^k$to be the transpose of the first right singular vector (the first column of **V**) times the first singular value (i.e., Δ_{11}). This updating process must be carried out for every column of the dictionary matrix **D**.

The process iterates between updates to the dictionary and updates to the coefficients (which assign sparse combinations of atoms to each word), until it reaches a predetermined stopping point. In our case, the process stops after either 10 iterations or the total reconstruction error falls below 1 times 10^{-6}, whichever happens first. The final result is a matrix of atom vectors **D** and a matrix of coefficients **X** that allow us to reconstruct each vocabulary word as a sparse linear combination of atoms. Conceptually, updating atoms in this way encourages distinct atoms; each time an atom is updated, the goal is to best account for all the variation in words' meanings that the other atoms do not already explain.

We note that our overall approach is extremely *modular*. While we use K-SVD to discretize the semantic space and identify topics, other dictionary learning algorithms (or even clustering algorithms, like k-means) can be used instead.[36] As long as this discretization returns a set of vectors in the embedding space, those vectors can be interpreted as topics (i.e., probability distributions over words) using the Latent Variable Model. They can also be mapped to the raw corpus using any sentence or document embedding technique to represent a stream of text as a context vector. The modular nature of DATM means that it can be *improved* as an overall strategy for text analysis following any innovation in these components, e.g., improvements to the Latent Variable Model (and SIF embeddings), to dictionary-learning algorithms, or to techniques that map context vectors to atoms.

DATM Model Quality and Selecting the Number of Topics

Measuring the quality of a topic model is important to validate that the model is learning human-interpretable topics and to aid in tuning model hyperparameters---most importantly, the number of topics (i.e., atom vectors). Evaluating topic model quality remains an open research area. Given the enormous number of possible models and topics within each model, we employ

[36] We conducted experiments using k-means. While performance on our corpus was comparable to K-SVD, we found that K-SVD produced interpretable topics more robustly across different corpora. We also note that the “theory of meaning” implicit in the K-SVD approach is more realistic: it views all words as a combination of basic semantic “building blocks” and finds those building blocks with that picture in mind. Using k-means implicitly assumes that the meaning of each word is best represented by the nearest cluster of word vectors, ignoring polysemy.

computational methods to evaluate topic model quality, in addition to human inspection and validation.

We trained candidate K-SVD models with the number of topics/atoms K ranging from 15 to 2000. We then used three metrics to evaluate model quality before selecting our final model. Our three metrics were: coherence, topic diversity (Dieng, Ruiz, and Blei 2020) and coverage (R^2). Together, these three metrics provide us with interpretable measures for: 1) how internally coherent topics are (coherence); 2) how distinctive topics are from each other (diversity); and 3) how well the topics explain or reconstruct the semantic space itself (coverage). Next, we explain each measure as implemented.

First, coherence is a commonly used family of metrics which attempts to measure the similarity of words within topics in a trained topic model (Aletras and Stevenson 2013; O'callaghan et al. 2015; Röder, Both, and Hinneburg 2015) To operationalize coherence, we first identified the top 25 word-vectors closest to an atom vector; for each atom vector, we then calculated the average pairwise cosine similarity between these closest word-vectors (Aletras and Stevenson 2013). Finally, we computed the average of these pairwise similarities across all atom vectors to arrive at an overall measure of coherence for the trained model. The coherence metric ranges from 0 to 1, where a value closer to 1 indicates higher average topic coherence (which typically corresponds to human interpretability of the topic, since the corresponding words are semantically similar).[37] As illustrated in Figure 8A, we found that models with fewer topics

[37] Hypothetically this coherence metric could range to -1, since cosine similarity between two word vectors in our word embedding may range from -1 to 1. In practice, word-vectors rarely have a negative cosine similarity. For clarity, we report this value as ranging from 0 to 1 in the

tended to produce slightly more coherent topics, but models were coherent across various numbers of topics.

Second, to measure how distinct topics are, we used an efficient and transparent metric: topic diversity (Arora et al. 2013; Dieng et al. 2020). To find diversity, we first identified the 25 word vectors closest to each atom vector in a model (with K atom vectors total). We then computed the proportion of these 25K “closest word instances” which are unique to a particular atom. If the top 25 words in every topic are unique, this measure will be 1.0, implying that the topics are very specific and distinct from one another. Topic diversity decreases as a larger number of words are repeated in the top 25 across multiple topics; it would reach its smallest value if the same 25 words were the “top 25” in all topics. As illustrated in Figure 8B, we found that models with fewer topics also tended to produce more distinct topics, and topic diversity dropped rapidly in models with more than approximately 225 topics.

While coherence and diversity favor a parsimonious topic model with few topics, it is nevertheless important that the model “explains” the space of possible meanings in the corpus. To capture this important aspect, we turned to our third metric: coverage. To measure how well the topics in a given model cover the semantic space, we computed the extent to which we could “reconstruct” the original semantic space using just the set of topics. As in k-means—which is in fact a special case of K-SVD (Aharon et al. 2006; Rubenstein et al. 2008) —the objective function of K-SVD minimizes the sum of squared errors between the original data and

main text (Mu, Bhat, and Viswanath 2017). This metric is well suited for topic modeling in embeddings, is efficient to compute, and correlates well with human judgement (Aletras and Stevenson 2013).

reconstructed data. Using the sum of squared errors and sum of squares total, we computed the proportion of the original variance explained by the topics (i.e., R^2) to measure how well a candidate set of topics explains the semantic space (we refer to the value for R^2 here as coverage). In contrast with topic diversity and coherence, coverage continues to increase in models with more topics, but the marginal gains from adding more topics reduce considerably around 225 topics in our data (Figure 8C).

We selected our final model to balance all three of our metrics for a good quality topic model. Coherence steadily decreased with more topics. Diversity dropped rapidly after around 225 topics. At first, coverage rapidly increased with more topics, but gained little after 225 topics. Thus, we selected a model with 225 topics as our final model. This model had a coherence of 0.59, a diversity of 0.93, and coverage of 0.63 (again, all metrics range from 0 to 1).[38]

[38] The final hyperparameter in the Discourse Atom Topic Model is the sparsity constraint T_0, which is the number of topics that a word in the embedding matrix is allowed to "load" on to (i.e., have a non-zero coefficient). The sparsity constraint must be between 1 (in which case K-SVD is identical to k-means) and the number of topics in the model. We follow Arora et al. (2018) in setting the sparsity constraint to 5. As they describe, if this sparsity constraint is not sufficiently low, then some of the coefficients must necessarily be small; this makes the corresponding components indistinguishable from noise (Arora et al. 2018). We empirically observed that models with more nonzeros have lower coherence and slightly less diversity, but higher coverage.

In other applications of K-SVD, Root Mean Square Error (RMSE) or the closely related Sum of Squared Errors (SSE) are used as metrics to select the number of elements (in our case, topics). To further inform our choice of the optimal number of topics, we plotted RMSE (or SSE) against the number of topics, and looked for the point at which adding more topics offers little reduction in SSE or RMSE. Both RMSE and SSE suggest that the optimal number of topics was approximately 250 (Figure 8D), quite close to the value selected by the procedure above balancing coherence, diversity, and coverage.

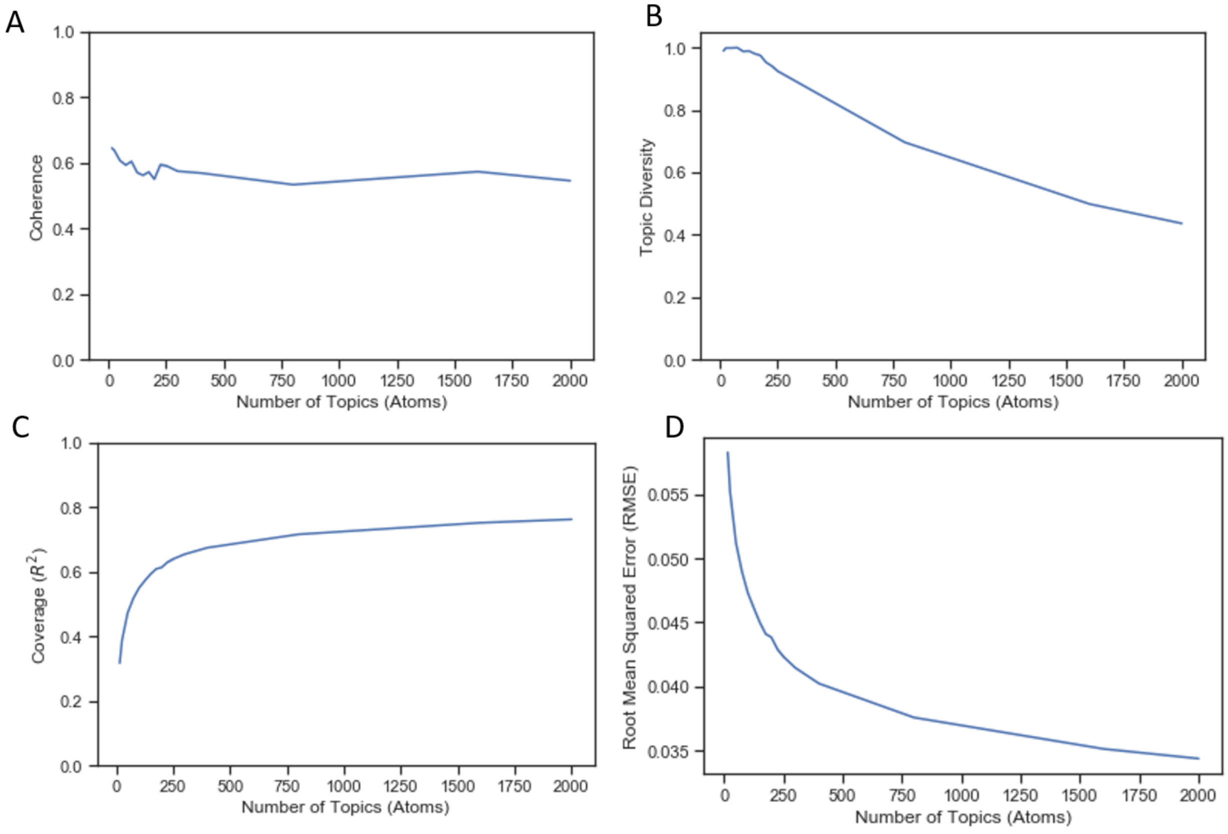

Figure 8. Figure A-D. Measures of Model Quality (Coherence, Diversity, Coverage, and RMSE) against the Number of Topics in the Discourse Atom Topic Model.

In Table 6 we list all the topics identified in our data using the Discourse Atom Topic Model. For each topic, we include our label (manually assigned) and the 10 most representative terms (from highest to lowest cosine similarity to the topic's atom vector).

Table 6. All 225 Topics identified using the Discourse Atom Topic Model

Topic Number	**Topic Label**	**Top 10 Most Representative Terms**
0	Taking (syntactic)	immediate_action, into_protective_custody, easy_way_out, sleeping_pill, antidepressents, pretty_hard, threat_seriously, too_many_pills, bath, own_life
1	Poisoning	mixed_drug_toxicity, ethylene_glycol_toxicity, ethylene_glycol_poisoning, carbon_monoxide_toxicity, acute_combined_drug_toxicity, mixed_drug_intoxication, cyanide_poisoning, combined_drug_toxicity, methadone_intoxication, natural_causes
2	Did not (syntactic)	know_what_happened, make_sense, understand_why, speak_english, hear_anything_else, regain_consciousness, know_why, express_suicidal_thoughts, acknowledge, recognize
3	Ligature around neck	ligature_encircling, encircling, ligature_furrow_encircling, encircles, encircled, fastened_around, tied_tightly_around, partially_encircling, wrapped_tightly_around, ligature_mark_around
4	Head	plastic_bag_covering, plastic_bag_tied_around, bag_covering, butted, blood_pooling_around, blood_pooled_around, fell_backwards_hitting, fell_backwards_striking, shaved, tilted
5	Drug supply	supplier, psychedelic, preferably, induced_psychosis, hallucinogenic, dependencies, sniffed, sells, abusers, users
6	Quotes 1	please_forgive_me, fucking, miss_you, i_hope, i_wish, sic, you_win, ya, whatever_happens, i_guess
7	From (syntactic)	official_sources, aside, borrowed_money, ranging, stealing_items, bad_odor_coming, ranged, refrain, polydrug_toxicity, russia
8	Victim number	1015, 1320, 1445, 1430, 1845, 1345, 1820, 1006, 2334, 1945
9	Knives	folding_pocket, butcher, serrated_edge, serrated_kitchen, switchblade,

		single_edged, fixed_blade, butterfly, serrated_blade, serrated
10	Specific times	0445, 2330_hours, 1030_hours, 1730_hours, 1445_hours, 1930_hours, 0920, 1330_hours, 0750, 2020_hours
11	Quotes 2	i_hate_my, poor_quality, falling_apart, enjoy_your, generally_unhappy, ruining, ruined, pointless, incompatible, boring
12	Neighborhood locations	fast_food_restaurant, retail_store, restaurant, mall, grocery_store, community_center, convenience_store, supermarket, shopping_center, strip_club
13	Race	american_indian_alaska_native, filipino, american_indian_alaskan_native, obese_caucasian, middle_eastern, indian, overweight_caucasian, asian_pacific_islander, latino, puerto_rican
14	Fumes	fumes, automobile_exhaust, inhaled, propane, natural_gas, gases, inhaling, intentionally_inhaling, exhaust_fumes, hydrogen_sulfide
15	Would (syntactic)	do_whatever, listen, press_charges, press_charges_against, follow, wait_until, join, assure, accept, sue
16	Education	grad, special_education, prestigious, vocational, doing_poorly, technical, rotc, graduate, freshman, junior
17	Around (syntactic)	goofing, noontime, moping, 11_30_pm, staggering, roaming, 7pm, 9_00_pm, doing_chores, started_throwing_things
18	Hedging	most_likely, entirely, definitely, definitively, likely, autoerotic_asphyxiation, natural, represents, strongly, purposeful
19	Up (syntactic)	really_messed, roughed, picks, hooking, sobering, doped, screwing, hose_hooked, messed, flared
20	Stains	stain, brownish, greenish, dark_brown, brown, reddish, dried, pink, blood_soaked, blood_stain
21	Gun shells and cartidges	shell, cartridge, spent_shell, live_round, spent_cartridge, 5_live_rounds, one_live_round, five_live_rounds, spent_round, shotgun_shell
22	Was (syntactic)	still_legally_married, legally_married, very_intelligent, unsure, rather, never_officially_diagnosed, nice_guy, generally_happy, unsure_if, un- confirmed
23	Body posture	wedged_between, hunched_over, resting_against, propped_up, slumped, leaning_against, propped_against, leaning_forward, lean- ing, slumped_forward
24	Common non-illicit medications 1	docusate, dicyclomine, valsartan, azithromycin, prochlorperazine, fenofibrate, levofloxacin, docusate_sodium, magnesium_oxide,

		metronidazole
25	Somatic symptoms	dizziness, nausea, headaches, fatigue, diarrhea, discomfort, severe_headaches, chills, stomach_pains, tremors
26	Psychiatric medications	dispensing, antianxiety, anti_seizure, anti_nausea, prescribed_psychotropic, anti_diarrhea, prescritpion, stockpiling, overtaking, mood_stabilizers
27	Multiple (syntactic)	aliases, inpatient_stays, psychiatric_admissions, jurisdictions, personalities, force_injuries, traumas, duis, abdominal_surgeries, knee_surgeries
28	Fall	backwards, cement, cinder_block, backward, stick, concrete, forward, brick, shovel, fell_backward
29	Locations in a house	master_bedroom_closet, bedroom, vicinity, living_room, laundry_room, backyard, sunroom, backroom, master_bedroom, front_yard
30	Clothing	robe, long_sleeve, sweater, boots, gloves, lanyard, drawstring, bathrobe, terry_cloth, rubber
31	Trouble keeping up	hygiene, skills, abilities, diminished, coping_skills, energy, decreasing, stability, cognitive, cognition
32	Not (syntactic)	yet_adjudicated, otherwise_defined, making_sense, otherwise_specified, surgical_candidate, eating_much, feeling_well_lately, mentally_stable, working_properly, strong_enough
33	Physical health conditions in older ages	hernia, diverticulitis, kidney_stones, colostomy, spinal_stenosis, replacements, mrsa, hip_replacement, bilateral_knee, ulcers
34	Grills and gas	charcoal, propane, portable, barbecue_grill, pan, grill, briquettes, burned_charcoal, heater, ash
35	Post-mortem examination	toxicological_examination, toxicological_screen, cat_scan, toxicologic_evaluation, rape_kit, toxicolgy, toxicological_analyses, post_mortem_examination, drug_screen, imaging
36	Clear forensic signs	displayed_obvious_signs, displaying_obvious_signs, certificatedeath_certificate_states, instantaneous, cert, certificate_lists, responding_medics_confirmed, sudden_cardiac, penalty, significant_conditions_contributing
37	Cause of death	hypoxic_ischemic_encephalopathy, anoxic_ischemic_encephalopathy, hypoxic_encephalopathy, hemorrhagic_shock, multisystem_organ_failure, hemothoraces, cerebral_disruption,

		cardiorespira- tory_arrest, hypovolemic_shock, respiratory_arrest
38	Drug user	heroin_abuser, methamphetamine_user, heroin_user, substance_abuser, crack_user, drug_user, meth_user, drug_abuser, crack_cocaine_user, heavily_consume_alcoholic
39	Drug paraphernalia	brown_substance, bottle_cap, plastic_baggie, metal_spoon, wrapper, glass_pipe, cotton_ball, grinder, cut_straw, foil
40	Physical health conditions	hyperlipidemia_hypertension, hyperthyroidism, vitamin_d_deficiency, osteoarthritis, peripheral_vascular_disease, obstructive_sleep_apnea, hypercholesterolemia, bph, peripheral_neuropathy, anemia
41	Miscellaneous observations and evi- dence	steak_knives, deformed_gray_metal_projectiles, capped_syringes, young_boys, teenage_boys, sealed_envelopes, failed_marriages, abor- tions, separate_incidents, young_children
42	Approximate time	approximately_30_minutes, approximately_15_minutes, nine_months, 30_mins, 15_mins, eight_months, 20_mins, 15_minutes, thirty_minutes, half_hour
43	Family and domestic	families, chests, wedding_anniversary, mouths, patrol_cars, wives, young_children, cellphones, cell_phones, wallets
44	Drinks	jack_daniels, nyquil, whisky, schnapps, brandy, champagne, bourbon, gin, tequila, rum
45	Physical aggression	tackled, lunged_toward, began_attacking, advanced_toward, attacked, slapped, intervened, shoved, lunged, pepper_sprayed
46	Miscellanous metrics	degrees_fahrenheit, bmi, wh_m, carboxyhemoglobin_saturation, mg_kg, saturation, ut_2014, mg_ml, mcg, co_level
47	Software and devices	gps, software, device, locator, tracking, tracker, heating, monitor, charger, aircraft
48	Paths of weapons into the body	overall_pathway, slightly_upwards, overall_path, temporal_scalp, pathway, downwards, red_purple, parietal_scalp, posteriorly, bullet_pathway
49	Suspicion and paranoia	conspiring_against, plotting_against, restraining_order_filed_against, belittled, please_forgive, making_fun, reminded, reminding, better_off_without, remind
50	Leftover alcohol and drug evidence	empty_liquor_bottles, liquor_bottles, prescription_pill_bottles, pill_bottles, loose_pills, empty_beer_bottles,

		empty_pill_containers, beer_cans, empty_beer_containers, insulin_syringes
51	Drug-related cognitive disturbances	groggy, lethargic, disoriented, incoherent, agitated, extremely_intoxicated, feeling_better, confused, acting_weird, fine
52	Canvassed	canvassed, nine_9mm_casings, canvased, drug_paraphrenalia, brushy, toured, fleeing, six_9mm_casings, 9mm_shell_casings, canvassing
53	Sedative and pain medications	phenergan, motrin, ultram, flexeril, endocet, lunesta, hydrocodone_apap, skelaxin, amitriptylin, norco
54	Altercation ensued	fight_ensued, gunfire_erupted, physical_altercation_ensued, another_individual, pistol_whipped, gunman, struggle_ensued, scuf- fle_ensued, suspect, intruders
55	Victim body parts	back_victim30, back_victim9, right_lateral_neck_victim22, back_victim48, chest_victim16, leg_victim37, chest_victim42, back_victim49, back_victim21, graze_type
56	Car crash	totaled, burst_into_flames, t_boned, jacked, totalled, backfiring, wrecking, impounded, intentionally_crashed, wrecked
57	Claims	advised, stated, added, explained, indicated, states, informed, claimed, relayed, said
58	Physical posture	crouching, silhouette, kneeling, northeast, crouched, walkway, platform, stagger, laying, leaning
59	Acute and multi-drug poisoning	acute_diphenhydramine, acute_multidrug, acute_opiate, acute_methanol, acute_salicylate, ethylene_glycol, multidrug, 11_difluoroethane, methanol, salicylate
60	Common non-illicit medications 2	testosterone, coumadin, ativan, antibiotics, estrogen, blood_thinner, dilantin, dilaudid, imodium, norco
61	Handguns	1911_45_caliber, smith_wesson_40_caliber, 380_semi_automatic, hi_point_45_caliber, beretta_9_mm, beretta_9mm, 9mm_hi_point, ruger_22_caliber, ruger_9mm, glock_40_caliber
62	Cleanliness	unkempt, messy, disorganized, cluttered, dirty, tidy, filthy, unclean, untidy, orderly
63	Wounds from physical impact	traumatic_brain, succumbs, superficial_sharp_force, craniofacial, conflagration, eventually_succumbed, massive_facial, facial, sustained_blunt_impact, non_survivable
64	Tried to (syntactic)	conceive, arouse, rouse, reassure, break_free, urinate, dissuade, restrain, assure, establish

65	Cognitive/emotional disturbances and decline	forgetful, irritable, needy, irate, insecure, introverted, moody, aggitated, argumentative, shaky
66	Fentanyl	fentanyl_4_fluoroisobutyryl, narcotic_fentanyl, des_propionyl_fentanyl_intoxication, furanyl_fentanyl_despropionyl_fentanyl, heroin_furanyl_fentanyl, fentanyl_despropionyl_fentanyl, fentanyl_4_fluoro, furanyl_fentanyl, despropionyl_fentanyl, cyclopropyl_fentanyl
67	Toxicology results	cocaine_benzoylecgonine, diazepam_nordiazepam_temazepam, ecgonine_ethyl_ester, nicotine_cotinine, cotinine_nicotine, methadone_eddp, diazepam_nordiazepam, cocaine_cocaethylene_benzoylecgonine, tramadol_o_desmethyltramadol, fluoxetine_norfluoxetine
68	Counseling	counselling, meetings, psychotherapy, outpatient_therapy, alcoholics_anonymous, therapy, counseling, grief_counseling, na_meetings, diversion
69	Cars	cargo_area, idling, drivers_side, windows_rolled_up, lone_occupant, t_boned, rear_hatch, rear_seat, front_seat, hatchback
70	Sums of money	50000, 30000, 20000, 8000, 40000, 100000, 7000, 6000, 3000, 5000
71	Native American	alaskan_native, alaska_native, natural_disease_process, prolonged_substance, soot_stippling, impending_criminal_legal, fabricating, pacific_islander, either, natural_diseases
72	Floor of building	3rd_story, 6th_floor, 2nd_story, 7th_floor, 5th_floor, 4th_floor, 12th_floor, 8th_floor, 10th_floor, 9th_floor
73	Writing materials	folder, notepad, legal_pad, manila_envelope, handwritten_letter, poem, book, poetry, spiral_notebook, folder_containing
74	Body dysfunction	aortic_aneurysm, aneurysm, enlarged_liver, enlarged_heart, abdominal_aortic_aneurysm, ulcer, abscess, umbilical_hernia, ovarian_cyst, enlarged_prostate
75	Psychiatric facilities	psychiatric_facility, psychiatric_ward, psychiatric_unit, psych_ward, mental_health_facility, inpatient_psychiatric_care, psych_unit, involuntarily_hospitalized, behavioral_health_facility, involuntarily
76	Partygoing and substances	partygoers, miscellaneous_items, topics, tested_substances, party_goers, factors, scenarios, ailments, prisoners, circumstanes
77	Chronic disease	atherosclerosis, cholelithiasis, pulmonary_anthracosis, nephrosclerosis, necrosis, fibrosis, aortic_atherosclerosis, hepatic_steatosis, hepatic, left_ventricular_hypertrophy

78	Things to jump or be pushed off	highway_overpass, freeway_overpass, bridge_overpass, railroad_bridge, railroad_trestle, train_trestle, viaduct, walkway, trestle, water_tower
79	Older ages	65, 68, 76, 71, 69, 67, 74, 66, 75, 55
80	Safety	unstated_reasons, safety_reasons, quite_awhile, sake, safe_keeping, actively_looking, safekeeping, mistaken, safety_purposes, quite_sometime
81	Idiopathic health conditions	idiopathic, chronic_fatigue_syndrome, myasthenia_gravis, scleroderma, anorexia_nervosa, diabetes_hypertension_hyperlipidemia, dysthymia, mental_retardation, bipolar_affective_disorder, irritable_bowel_syndrome
82	Long decomposed body	badly_decomposed, partially_skeletonized, frozen_solid, severely_decomposed, charred, partially_decomposed, mummified, heavily_decomposed, dismembered, decaying
83	Death records	examiner_investigator_mei, records_reflect, examiner_opined, investigator_2016_74, professionals, marijuana_card, investigator_2015, investigator_2014, examiners_office, records_show
84	Pawned	jewelry, pawned, fake, pawn, concealed, owning, merchandise, traded, debit_card, valuables
85	Propped open	propped_open, slightly_ajar, slightly_open, cracked_open, ajar, rear_sliding_glass, pried_open, barricading, top_hinge, wide_open
86	Extreme amounts	weighed_230_pounds, weighed_240_pounds, weighed_200_pounds, weighed_250_pounds, 45am, yrs_ago, daughters_ages, 15am, inches_deep, inch_laceration
87	Doing something against one's will	carjack, abduct, dissuade, evade, intubate, persuade, administer_first_aid, reviving, render_first_aid, reconcilliation
88	Prior encounters with the legal system	protective_order, temporary_restraining_order, restraining_order, protection_order, bench_warrant, summons, traffic_ticket, dui_charge, citation, protective_order_against
89	Military guns and weapons	springfield_armory, springfield, hi_point, bersa, walther, highpoint, sturm_ruger, sig_saur, kimber, ruger_lcp

90	Chronic mental instability and substance use	depresion, alcoholism, problematic_alcohol_use, paranoid_schizophrenia, manic_depression, depresssion, chronic_alcoholism, substance_abuse, intravenous_drug_abuse, psychiatric_illness
91	Acting strangely lately	drinking_heavily, drinking_excessively, acting_strangely, exchanging_text_messages, acting_strange_lately, sending_text_messages, acting_paranoid, acting_differently, acting_erratically, acting_strange
92	Unknown details	exact_timeframe,timeline, marital_status, exact_time_frame, if_contributing_condition, timeframe, timeframes, exact_timing, if_contributing_factor, downtime
93	Messages	deleted, unread, unsent, listened, went_unanswered, pinging, sends, forwarded, draft, wireless
94	Tying rope-like materials	tied, fastened, tightened, twisted, tying, wrap, knotted, loosened, looped_around, draped
95	Organ failure	dvt, paroxysmal,venous_insufficiency, elevated_liver_enzymes, esophageal_reflux, pvd, hypertension_atrial_fibrillation, iron_deficiency_anemia, hypo, hcc
96	Hotlines and government institutions	crisis_hotline, crisis_line, law_enforcements, poison_control, police, sheriff_deputies, dispatch, alarm_company, sheriffs_office, law_enforcement
97	As (syntactic)	whow, made_statements_such, categorized, making_statements_such, phrases_such,precaution,wells,best_certified,serves, train_got_closer
98	Over (syntactic)	seas, despondant, despondence, financial_matters, counter_cold_medicine, court_battle, become_increasingly_despondent, counter_sleep_aids, counter_sleeping_pills, hovering
99	Missing, runaway, and endangered	runaway, missing_endangered_person, whow, precaution, teenager, making_statements_such, made_statements_such, categorized, miss- ing_endangered, dispute_words_exchanged
100	Out of doors	landscape, plumbing, fish, roofing, heating, construction, catering, ski, sports, fields
101	Case number 1	1174, 1170, 328, 907, 949, 491, 617, 766, 486, 708
102	Games	games, game, video_game, volleyball, computer_games, dominoes, beer_pong, football_game, tennis, chess
103	Everything seemed fine	fell_asleep, everything_seemed_fine, seemed_fine, wakes_up,

		ran_errands, ate_breakfast, watched_television, woke_up, ate_dinner, woke
104	Blood alcohol level	096, 179, 218, 0_08, 247, 0_02, 0_01, 246, g_100_ml, 390
105	News and official reports	news_reports, news_articles, court_records, press_release, da_press_release, newspaper_reports, court_documents, da, media_reports, newspaper_articles
106	First aid and CPR	administered_cpr, fire_dept, initiated_cpr, first_aid, continued_cpr_until, began_cpr_until, rescue_personnel, basic_life_support, initiated_cpr_until, swat
107	Lacerations	lacs, fresh_cuts, elbows, superficial_incisions, thighs, slicing, punctures, slash_marks, shins, cut_marks
108	Transfer to medical institution	life_flighted, trasnported, air_lifted, medflighted, readmitted, transfered, upon_admission, despite_resuscitative_measures, mental_ward, airlifted
109	Involved parties	involved_party2, involved_party3, involved_party1, related_party1, concerned_party2, witness3, witness4, reporting_party, related_party2, involved_party
110	Young adults	youths, young_men, juveniles, individuals, teenagers, women, men, students, adults, parties
111	Characteristics of suspects	uid, tank_top, hoody, hooded_sweatshirt, tar_heroin, dreads, velvet, sheep, tarry, ski_mask
112	Intentions and desires	intending, willing, must, wanted, wants, suppose, didn_t_want, intended, pretending, wanting
113	Quotes 3	happiness, i_truly, sic, hate, i_wish, life_sucks, i_hope, god_bless, i_hate, soul
114	Quotes 4	loves, hated, hates, loved, will_miss, proud, worthless, felt_like, hoped, disappointment
115	Employer-related	apparant, excellent_employee, occasional_drinker, illegal_immigrant, abandoned_warehouse, illegal_alien, unnamed_citizen, avid_hunter, off_duty_firefighter, addictive_personality
116	Body parts	lower_lung_lobe, ulna, upper_lung_lobe, subclavian, psoas_muscle, humerus, 10th_rib, hemidiaphragm, subclavian_artery, axilla
117	Gangs and criminal	gang, rival_gang, bloods, crips, gang_activity, drug_trade, crips_gang,

	networks	rival, rival_gang_members, revenge
118	Large amounts	delayed_effects, lots, hundreds, bunch, large_number, variety, excessive_amounts, substantial_amount, large_quantity, large_amount
119	Absence of signs and information	circumstantial_info, circumstancial_info, narrative_available, foul_play_suspected, detectable_pulse, ages_given, success, fixed_address, brain_activity, elaboration
120	Specific outdoor locations	drainage_ditch, ravine, grassy_area, pasture, vacant_lot, wooded_area, pond, field, public_park, natural_area
121	Chairs	lounge_chair, folding_chair, recliner_chair, plastic_lawn_chair, lawn_chair, rocking_chair, reclining_chair, cross_legged, porch_swing, chair
122	Went to do something	their_separate_ways, grocery_shopping, lie_down, golfing, buy_cigarettes, bowling, bike_ride, lay_down, do_laundry, cool_off
123	Filth and disarray	walls, human_feces, blood_smears, smears, stacked, scattered_throughout, clutter, broken_glass, bloody_footprints, dirty_dishes
124	Projectile	metal_projectile, jacketed_bullet, jacketed_projectile, lead_bullet, gray_metal_projectile, copper_colored_projectile, jacket_fragment, copper_jacket, copper_jacketed_bullet, copper_jacketed_projectile
125	On (syntactic)	weekly_basis, may_X_20XX, numerous_occasions, many_occasions, depending, regular_basis, daily_basis, private_property, occassion, operating_table
126	Housekeeping	housekeeping, maid, desk_clerk, cleaning_staff, housekeeping_staff, housekeeper, hotel_staff, maintenance_staff, motel_staff, hotel_management
127	Recent social interactions	visited, texted, emailed, text_messaged, spoken, passed_away, confided, sent_text_messages, expressed_suicidal_ideations, split_up
128	Mass murder	children_ages, unspent_rounds, daughters_ages, blocks_away, inch_barrel, 380_caliber_casings, wheelers, wheeling, deformed_gray_metal_projectiles, 40_caliber_cartridge_casings
129	Harassing	harassed, stalked, harrassed, unfaithful, consoled, taken_advantage, disrespectful, remodeled, held_hostage, victimized
130	Case number 2	2205, 2111, 0002, 1526, 2038, 1113, 1853, 2039, 1719, 2134
131	Out of air	smothering, oxygen_displacement, helium_gas_inhalation, helium_inhalation, oxygen_exclusion, oxygen_deprivation, neck_compression, upper_airway_obstruction, vitiated_atmosphere, oxygen_depletion

132	Medical professionals	neurologist, mental_health_provider, psychologist, specialist, mental_health_professional, counselor, psychotherapist, therapist, psychiatrist, pain_management_doctor
133	Discoloration (graphic)	discoloration, purple, purge, bluish, fly_eggs, discolored, bright_red, coloring, blistering, lips
134	Rifling through car	rifled, coursing, sunroof, rear_passenger_window, passenger_side_window, rear_window, sliding_glass_door, windshield, window_blinds, backdoor
135	Enforcement agency	kentucky, highly_decomposed, southern, bureau, bordering, highway_patrol, file_prince_george, new_york, council, corrections
136	Problematic arrest	apparant, illegal_immigrant, extensive_criminal_record, ak_47_rifle, abandoned_warehouse, excellent_employee, occasional_drinker, exacto_knife, upcoming_court_appearance, extention_cord
137	Finances	bonds, funds, credit, 6000, 401k, checkbook, stocks, 8000, 20000, bank_card
138	Acting strangely, drug-related	act_strange, freaking_out, poking, act_strangely, seizing, snore, convulsing, physically_fighting, harass, fussing
139	Died in hospital	extubated, intubated, asystolic, never_regained_consciousness, declared_brain_dead, condition_declined, rhythm, asystole, surgical_intensive_care_unit, pulseless_electrical_activity
140	Turmoil	turmoil, discord, strife, friction, difficulties, insecurity, instability, interpersonal, disagreements, romantic
141	Rifles and shotguns	savage_arms, bolt_action, savage, pump_action, mossberg, stoeger, stevens, 410, remington, 12_gauge_winchester
142	Things to hang from	ceiling_rafter, ceiling_rafters,metal_beam, stair_railing, spiral_staircase, ceiling_beam, light_fixture, ceiling_joist, support_beam, bedpost
143	Alcohol	malt_liquor, 24_oz, budweiser, bud_light, 40oz, 40_ounce, 24oz, 16_oz, miniature, miller_lite
144	Military	reserves, air_force, us_army, army, national_guard, armed_forces, afghanistan, active_duty, navy, tour
145	Sitting at	sofa, dining_room_table, dining_table, kitchen_table, park_bench, couch, carpeted_floor, kitchen_counter

		television_stand, weekly_basis
146	Proof	proven, ascertained, fingerprinted, proved, established, reviewed, deemed, processed, swabbed, photographed
147	Uncertainty	unlikely, uncertain_whether, too_late, looks_like, unclear_why, seems, speculated, looked_like, rumored, raining
148	Escalation	becoming_increasingly, becoming_more, become_more, increasingly, noticeably, notably, become_increasingly, profoundly, grown_increasingly, generally
149	Illegal narcotics	crystal_meth, crystal_methamphetamine, illegal_narcotics, spice, recreational_drugs, synthetic_marijuana, meth,crack_cocaine, heroin_e, herion
150	Mental illnesses	ocd, obsessive_compulsive_disorder, bipolar_disorders, borderline_personality_disorder, generalized_anxiety_disorder, mania, schizoaffective_disorder, oppositional_defiant_disorder, delusional_disorder, borderline_personality
151	Amounts of substances	045, 050, 040, 060, 026, 025 cocaine_cocaethylene_benzoylecgonine, 018, 083, 047
152	Gun actions	always_carried, cleaning_equipment, enthusiast, dry_firing, battle_ensued, cleaning_kit, 357_cal, malfunctioned, reloaded, acciden- tally_discharged
153	Institutional involvement	mandated, sponsored, technical, administrative, mandatory, representative, educational, assistant, vocational, restoration
154	Limited information and evidence	supplemental, supplementary, redacted, very_limited, concludes, conflicting, contains_little_useful, extremely_limited, gives, extremely_brief
155	Brain trauma	hemorrhages, basilar_skull_fractures, cerebral_contusions, ecchymoses, hemorrhaging, subscalpular, bilateral_periorbital, sub- galeal_hemorrhage, subdural_hemorrhages, fractures
156	Incarceration	cell_block, bunk, top_bunk, segregated, cell, segregation_unit, maximum_security, solitary_confinement, correctional_institute, guards
157	Letters	multi_page, computer_generated, farewell, hand_written, titled, typed, entitled, outlining, sticky, detailing
158	States and countries	california, florida, arizona, colorado, mexico, texas, united_states, oklahoma, minnesota, ny
159	Observers	passers, passer, drowning_complicated, attendant_doctor, ingesting_pills, pass_surgery, complicated, intoxication_complicated, teleme- try, county_coroner
160	Drinking	consumed_large_quantities, occasionally_drank,consumed_large_amounts,consuming_large_amounts, smoked_cigarettes_drank, consumes, smelled_strongly, consuming_large_quantities, smelled_heavily, consume_large_amounts
161	Past suicidal attempts and	commited, contemplated, commiting, commits, x74, comitted, contemplating, would_often_threaten, contemplate, committ

	ideation	
162	Recluse behavior and illness	recluse, heavy_drinker, very_ill, chronic_alcoholic, bedridden, reclusive, recovering_alcoholic, forgetful, mentally_unstable, legally_blind
163	Surveillance	surveillance_camera, surveillance_video, security_camera, video_surveillance, footage, surveillance_footage, security_cameras, convience, surveillance_cameras, security_footage
164	Cognitive actions	deliberation, awaking, pleading_guilty, reviewing, learning, shorty, thorough_investigation, noticing, gaining_access, gaining_entrance
165	Appointments	twenty, follow_up_appointment, scheduled_appointment, appoint, appointment_scheduled, seventeen, thirty, forty, appointment, surgi- cal_procedure
166	Numbers	six, seven, eight, four, five, 6, twelve, 00_buck, 4, nine
167	Dead body position	face_down, facedown, fetal_position, non_responsive, supine, hunched_over, lifeless, fully_clothed, nude, sitting_upright
168	Self-injurious behavior	starving, joked_about_killing, isolating, suspending, may_have_harmed, asphyxiating, defended, intentionally_suffocating, voluntarily_committed, suspend
169	Older and middle-aged life stressors	digestive, gender_identity, economic, experiencing_intimate_partner, urinary_tract, having_martial, circulatory, pending_legal, impulse_control, gastro_intestinal
170	Police-related	reponded, tracked_down, evading, kansas_state, headquarters, arresting, baltimore_city, canvassed, eluding, uniformed
171	Sneaky movements	snuck, darted, blacking, neatly_laid, lashed, lashing, spaced, hangs,nodding, chickened
172	Extra victim	extra_victim3_extra, extra_victim5, 453, 1173, 492, 447, 808, 301, 749, 988
173	Seemed like	apprehensive, fifteen_minutes_later, seemed_excited, nobody_cares, very_secretive, ten_minutes_later, 150_feet, nobody_cared, half_hour_later, items_strewn
174	Digital communication	snapchat, facebook_message, tm, snap_chat, tms, message, emails,text, text_message, disturbing_text_message
175	Metrics	qty, g_respectively,bmi, wh_m, 06, mcg, 90_remaining, merged_into_ky_2017, olds, mg_kg
176	People	guy, girl, young_girl, liar, boy, young_man, kid, woman, gangster, mutual_friend
177	Forensic analyses	studies, analysis, antemortem, analyses, ante_mortem, inconclusive, non_contributory, prolonged_hospitalization, laboratory, post- mortem_toxicological
178	Waning engagement	dozed, dozing, tapered, brushed, label_torn, taper, grid, fend, cooled, tapering

179	With (syntactic)	having_difficulty_dealing, having_trouble_dealing, medics_pronouncing,coupled, dealt, certainty, obsessed, strained_relationships, interfering, coincided
180	Suicidal	self_destructive, suidical, overt, sucidal, subtle, passive, suicidial, suicial, self_harm, fatalistic
181	Living situation	retirement_community,transitional_housing, boarding_house, rooming_house, senior_living_facility, low_income, transitional, halfway_house, transient_lifestyle, independent_living
182	Preparation for death	disposal, deeds, prepaid_funeral, burial, worldly, miscellaneous, pawning, distributed, giving_away, pre_paid
183	Right side of body	right_temporal_bone, proximal, right_frontal, frontal_bone, inferior,right_temporal, basilar_skull, overlying, parietal_scalp, skull_base
184	Descriptions of time sequence	shortly_after, after, shortly_thereafter, upon, eventually, shortly_before, subsequently, af, upon_arrival, moments_later
185	Personality and behavior	weird, grumpy, cheerful, loopy, moody, irritable, shy, upbeat, quiet,pleasant
186	In (syntactic)	addition, 1994, 1970, 1995, 2017_1078, late_afternoon_hours, meantime, 2002, 1999, 1997
187	Month	april, september, june, february, october, august, november, march, july, january
188	Observations	apparant, abandoned_warehouse, occasional_drinker, excellent_employee, illegal_immigrant, empty_wine_glass, unsent_text_message, addictive_personality, unnamed_citizen, extensive_criminal_record
189	Making (syntactic)	amends, bad_decisions, inappropriate_comments, similar_comments, advances_towards, advances_toward, bad_choices, poor_decisions, similar_statements, statments
190	Relationships	strained_relationships, interacted, having_sexual_relations, sexual_relationship, flirting, certainty, sexual_relations, beef, interacting, conversing
191	Forced entry	single_wide_mobile, split_level, forcibly_entered, invaded, burglarized, trashed, ranch_style, complete_disarray, condemned, quarrel_location
192	Rope materials	nylon_strap, bungee_cord, tow_strap, nylon_belt, nylon_cord, ratchet_strap, yellow_nylon_rope, shoestring, tow_rope, necktie
193	Belongings	belongings, permission, pension, loan, social_security_benefits, possessions, stuff, groceries, gifts, food_stamps
194	Decomposed body (graphic)	bloated, bloating, mummification, maggots, skin_slippage, marbling, insect_activity, blistering, maggot_activity, partially_mummified

195	Loud noises	loud_crash, loud_thump, pop_noise, loud_boom, loud_thud, loud_sound, boom, popping_sound, loud_pop_sound, muffled_bang
196	Scheduled event	doctors_appointment, court_hearing, court_appearance, meal, counseling_session, christmas, holiday, thanksgiving, follow_up_appointment, valentine
197	Children	biological, eldest, newborn, grandchild, foster, youngest, toddler, fathered, molesting, sexually_abusing
198	Frequency of time 2	may_14th_2015, numerous_occasions, many_occasions, weekly_basis, several_occasions, operating_table, 23rd_2015, depending, regular_basis, 11th_2015
199	Hostile interactions	home_invasion_robbery, card_game, drug_transaction, gunfight, shootout, scuffle, hostage_situation, verbal_exchange, confrontation, brawl
200	Substance dependency	nicotine_dependence, generalized_anxiety, narcotic_dependence, recurrent_major, opiate_dependence, episodic, hypothyroid, alcholism, bulemia, recurring
201	Illegals	illegal_immigrant, ongoing_issue, abandoned_warehouse, absentia_case, unnamed_citizen, armed_security_guard, illegal_alien, active_restraining_order, extensive_criminal_record, outstanding_felony_warrant
202	Decay	whose_skeletal_remains, spouses, only_ones, alcoholics, heroin_addicts, recovering_alcoholics, heavy_drinkers, manners, strangers, though_its_contents
203	Crimes	grand_larceny, felony_menacing, criminal_mischief, misdemeanor, simple_assault, criminal_trespass, reckless_endangerment, felony_theft, assault_battery, retail_theft
204	Justification	ruled_justifiable, remains_unsolved, 558, gang_motivated, pedestrian_vs_train, road_rage, random_violence, justifiable_self_defense, 3289, considered_justifiable
205	Traffic	tractor_trailer, semi_tractor_trailer, semi_truck, into_oncoming_traffic, dump_truck, westbound, eastbound, swerving, median, southbound
206	Treatment	resistant, began_administering, intensive, ect, life_saving, receives, currently_undergoing, requiring, undergoing_radiation, discontinued
207	Interpersonal violence	extra_victimuspect, prime, pn, coh, physical_altercation_ensued_between, began_attacking, claiming_self_defense, person5, gh, hep
208	Water	shore_line, boat_ramp, waterway, waters, fresh_water, dam, basin, partially_submerged, harbor, downstream

209	Cancers	lymph_nodes, malignant, lymph_node, metastases, lesion, adenocarcinoma, cancerous, pancreas, metastatic, metastasized
210	Financial problems	having_financial_difficulties, having_financial_problems, experiencing_financial_problems, experiencing_financial_difficulties, financial_problems, struggling_financially, financial_difficulties, having_financial_troubles, financial_issues, bankruptcy
211	Short-range weapons (knives and handguns)	steak_knife, 380_caliber_handgun, kitchen_knife, 40_caliber_handgun, butcher_knife, 38_caliber_handgun, 45_caliber_handgun, 9mm_handgun, 32_caliber_revolver, 357_caliber_handgun
212	Expressed suicidal ideations	expressed_suicidal_ideations, voiced_suicidal_ideations, knee_surgery, nervous_breakdown, expressed_suicidal_ideation, expressed_suicidal_thoughts, pacemaker_installed, miscarriage, hip_surgery, verbalized_suicidal_ideations
213	Drug concentrations	concentrations, levels, concentration, toxic_levels, toxic, lethal_levels, therapeutic_levels, elevated_levels, toxic_level, production
214	Tubes	tubing_connected, tube_attached, tubing_leading, tube_connected, tubing, tubing_attached, plastic_tubing, helium_gas_tank, nitro- gen_tank, plastic_tube
215	Sexual body appendages	penis, right_arm, arm, right_leg, genitals, throat, breasts, right_ankle, fingers, inner_thigh
216	Worker	contractor, worker, co_worker, coworker, landscaper, construction_worker, customer, employee,maintenance_worker, delivery_person
217	Causal language	sparked, preceded, triggered, precipitated, led, prompted, culminated, may_have_contributed, occured, completely_unexpected
218	Containers	suitcase, garbage_can, cardboard_box, dresser_drawer,plastic_container, beer_can, briefcase, drawer, bible, pill_bottle
219	Rural outdoor areas	wooded, densely, mountain, farmer, rural, hiker, forested, muddy, swamp, picnic_area
220	Body fluids	splatters, oozing, urine_samples, saturated, coagulated, congealed, exuded, large_puddle, smeared, thinning
221	Foul odor	foul_odor, foul_smell, bad_odor, bad_smell, foul_odor_coming, mail_piling_up, strong_foul_odor, foul_order, foul_smell_coming, strong_odor_coming
222	Semi-auto pistol manufacturers	springfield_armory, springfield_arms, davis_industries, bersa, hi_point, springfield, walther, kimber, sig_sauer, cobra
223	Directions	west, south, east, north, south_side, west_bound, westbound, northbound, avenue, southbound
224	Family members	mother, grandmother, father, sister, niece, aunt, stepfather,

		brother, fiance, best_friend
Note: Most representative terms are listed in order of highest to lowest cosine similarity to the topic's vector. Misspellings which were not caught in preprocessing are retained here. Topics which are observed to be syntactic are denoted with (syntatic) in the topic label. Topics which may be graphic are denoted with (graphic) in the topic label. One term is modified in this table to "may_X_20XX" to retain anonymity.		

Comparing DATM to Other Topic Models

Topic modeling is a core method in text analysis. It is therefore unsurprising that a plethora of specific approaches exist. These include non-negative matrix factorization, joint-stochastic matrix factorization, matrix rectification, Sparse Additive Generate Models (Eisenstein, Ahmed, and Xing 2011), anchor-based topic modeling (Arora et al. 2013; Arora, Ge, and Moitra 2012), replicated softmax (Hinton and Salakhutdinov 2009), Latent Semantic Analysis (Landauer, Foltz, and Laham 1998) and its variants, and most notably a wide variety of latent Dirichlet allocation (LDA) topic models and implementations (Blei 2012; McCallum 2002; Řehůřek and Sojka 2010)

DATM differs from this prior work in several ways. Most crucially, DATM differs from the majority of topic models (like LDA) in that it integrates topic modeling with word embedding, capitalizing on the distinct capabilities of each of these core methods. Several other recent topic models also aim to combine word embedding and topic modeling, but do so in ways that are quite different from DATM. Many of these models simply use information about words derived from word embeddings, such as word similarity, to inform the construction of the topic model (e.g., Das, Zaheer, and Dyer 2015; Petterson et al. 2010; Xie, Yang, and Xing 2015; Zhao, Du, and Buntine 2017). In contrast, DATM directly represents a topic in an embedding space (as does Dieng et al. 2020)

DATM is also distinct from prior topic modeling because it leverages a generative model that connects word embedding itself to observed text (i.e., the Latent Variable Model) (Arora, Li, Liang, et al. 2016; Arora et al. 2017). It does not rely on LDA (or any variant of LDA) for a generative model. This is a critical distinction with Dieng et. al. (2020), which represents documents as mixtures of topics, as in traditional LDA. DATM, by contrast, represents documents as a sequence of topics, based on a discretization of the inferred context vector position. This sequence can be converted into a distribution over topics, or into a binary presence/absence representation as we do here (as in Hinton and Salakhutdinov 2009). As illustrated in this paper, because DATM topics exist directly in the embedding space, researchers can easily extend methods commonly used to work with words in word embeddings (e.g., extracting biases and cultural dimensions, like gender) to work with DATM topics.

Another crucial difference between DATM and other topic models is that its input is a semantic space derived from the corpus (i.e., a word embedding trained on the corpus). In contrast, the input to other topic models is usually document-level word counts, e.g., a document-term matrix. To code documents with topics, DATM maps topics onto the sequence of local, inferred context vectors (a "trajectory") that represents each document in semantic space; it thus distills each document into a sequence of topics. Ignoring order, this sequence can be converted into a distribution, or even a vector of binary presence/absence indicators. DATM can be thought of as following a "bottom-up" approach to inferring the topics in a document; it is fundamentally different from the traditional, "top-down" approach to topic modeling, which includes further assumptions about the role of topics in the text-generation process.

DATM has several practical advantages compared to many prior topic models. It is robust to stopwords, domain specific vocabulary, and can be used on documents of varying lengths. It can also yield highly interpretable and coherent topics, as we illustrate in this paper. As we show next, the topics identified by DATM are qualitatively different from those picked up with LDA topic models (the mainstream approach in computational social science). We emphasize, however, that the "ideal" topic model for a particular use-case will depend on the researcher's data, theoretical assumptions, and research questions.

LDA Topic Modeling on NVDRS Narratives

Here we provide sample topics generated on our data using one of the most popular topic models: LDA topic modeling (Table 7). Our goal is not to show that the Discourse Atom Topic Model necessarily works better than any other topic model in general. Rather, our goal is to highlight that our model and LDA pick up qualitatively different topical structures in our data and DATM can answer different questions compared to traditional topic modeling approaches.

To train our LDA topic models, we used a Python wrapper (Řehůřek and Sojka 2010) for the MALLET implementation of LDA topic modeling (McCallum 2002), after observing that this implementation offered substantially more interpretable topics than the default implementation in Python using the Gensim package (Řehůřek and Sojka 2010). For instance, in an LDA model trained with 225 topics using the default implementation, the most five probable terms for one topic (topic 214, selected at random) include: "hispanic," "homeless," "wood," "decomposing," and "inflicted." The five most probable terms for another randomly chosen topic (topic 154) include: "seen_alive," "last," "initiated," "doorway," and "letters." For reference, the overall model had a coherence of 0.11 and a topic diversity of 0.69. As a second example, in an LDA model trained with 100 topics, the most five probable terms for one topic (topic 60, selected at random) include: “garage,” “nature,” “dog,” “fatal_injury,” and “contents.” This overall model scored similarly on coherence and diversity (0.12 and 0.66, respectively).

We initially tried using the exact same vocabulary as we used for our Discourse Atom Topic Model. However, the resulting topics were uninterpretable. They contained many stopwords and words that are very common in our data and thus lose meaning (e.g., “the” and “victim”). LDA models require careful pre-processing that is specific to the corpus, and often are not robust to stopwords (Boyd-Graber et al. 2014; Schofield, Maans Magnusson, and Mimno 2017; Schofield, Mans Magnusson, and Mimno 2017), unlike the Discourse Atom Topic Model. Thus, for LDA topic modeling, we removed standard stopwords using a list from the nltk package in Python (we retained gender pronouns, however, even though these are considered stopwords in the nltk list). We also removed words that occurred in more than 75% of the documents (ubiquitous words), or fewer than 15 times total in the corpus (very rare words).

To select the best LDA model, we trained 11 LDA models with varying values of K (i.e., topics): 15, 25, 50, 100, 150, 175, 200, 225, 250, 400, and 800. We selected our final LDA topic model among these using coherence and diversity metrics, described in the main paper; coverage does not apply to LDA topic models, since it has to do with the ability of topic atoms to reconstruct an embedding space. To compute coherence and diversity, we selected the “top 25” words for each topic by considering their probability given the topic. In our LDA topic models, coherence has minimal gains after 100 topics (when coherence is 0.18); it then drops with more than 250 topics. Topic diversity begins at 0.66 (at 15 topics, the smallest number of topics we considered) but rapidly diminishes (e.g., by 250 topics the topic diversity is 0.41). Using the elbow method, we selected a final LDA model with 100 topics to balance both coherence and diversity.

Table 7. All 100 Topics identified using LDA Topic Modeling

Topic Number	Top 10 Most Representative Terms
0	stating, phone, text_message, received, text, i_m, stated, left, message, text_messages
1	bedroom, bed, floor, lying, deceased, living_room, kitchen, room, discovered, face_down
2	report, death, nothing_further, police, alive, pronounced, medical, unresponsive, incident, manner
3	heroin, drug, cocaine, history, drugs, unresponsive, drug_abuse, methamphetamine, drug_paraphernalia, marijuana
4	hanging, neck, rope, suicide, ligature, closet, cut, belt, tree, basement
5	suicide, death, history, note, report, medical, law_enforcement, manner, called_911, medications
6	depressed, recently, job, lost, due, suicide, depression, family, problems, recent
7	night, morning, bed, hours, evening, sleep, sleeping, couch, called_911, woke_up
8	blood, observed, located, noted, mouth, body, appeared, feet, top, side
9	prescription, pills, overdose, oxycodone, empty, medication, bottle, medications, bottles, filled
10	ago, months, 3, years, 5, 4, 2, weeks, 6, days
11	death, states, police, pronounced, scene, method, medical, 58, white, alcohol
12	girlfriend, roommate, exgirlfriend, relationship, told, broken_up, night, recently, roommates, kill_himself
13	medical, emergency, service, scene, pronounced, arrived, declared, 1814, plethora, ev
14	details, died, unspecified, time, firearm, place, residence, body, age, location
15	suspect, homicide, murder, killed, charged, arrested, fled, altercation, kill, killing
16	gun, shot, head, pulled, put, trigger, firearm, guns, shooting, fired
17	mention, attempts, note, diagnosis, depressed_mood, mental_health, information, intentional, given_regarding, threats
18	gunshot, shot, weapon, wounds, chest, recovered, handgun, fired, shooting, multiple
19	wife, divorce, separated, estranged, told, children, problems, kill_himself, marriage, spouse
20	stated, witness1, time, told, spoke, witness2, left, located, back, mentioned
21	pain, back, suffered, chronic, surgery, medications, due, doctor, years, chronic_pain
22	suspect1, suspect2, suspects, hispanic, shot, robbery, murder, charged, arrested, suspect3
23	death, unresponsive, undetermined, manner, history, pronounced, intoxication, drug, signs, methadone
24	residence, white, back, house, 46, 54, yard, backyard, 60, died

25	hospital, transported, died, admitted, injury, local, days, staff, complications, transferred
26	reported, incident, reports, died, family, time, occurred, decedent, believed, date
27	black, unknown, shot, circumstances, 19, age, relationship, 18, suspects, age_race
28	wound, head, gunshot, self_inflicted, revolver, handgun, weapon, 38_caliber, hand, 22_caliber
29	cancer, white, diagnosed, health, care, recently, dementia, 69, doctor, 72
30	work, day, morning, show_up, co_worker, employer, worked, failed, working, called
31	suicide, attempted, depression, history, attempts, attempt, past, previous, overdose, suicidal_ideations
32	father, parents, school, family, 16, 17, 15, 14, 18, grandfather
33	unresponsive, responded, scene, alive, report, pronounced, death, medical, physical, reported
34	door, room, locked, open, hotel, entered, opened, motel, checked, check
35	store, parking_lot, business, owner, back, building, restaurant, office, local, employee
36	house, fire, inside, body, set, home, trailer, church, due, property
37	information, report, apparent, regarding_circumstances, nature, suicidal, dead, toxicology, positive, suffering
38	multiple, neck, chest, knife, stabbed, stab_wounds, injuries, stab, times, abdomen
39	apartment, neighbor, neighbors, heard, window, floor, apartment_complex, door, resident, inside
40	dead, reportedly, subject, white, body, confirmed, suicidal, scene, recently, death
41	alcohol, abuse, history, problem, 44, drugs, 36, 37, scene, previous
42	death, information, time, ruled, police, toxicology, report, additional, responded, determined
43	brother, sister, family, law, family_members, home, family_member, nephew, day, house
44	money, state, pay, move, living, years, financial_problems, rent, property, bills
45	told, asked, wanted, back, leave, needed, thought, talk, started, refused
46	white, death, due, 28, 35, gunshot, wound, 56, female, homocide
47	home, returned, left, day, earlier, find, return, work, morning, returning
48	son, white, home, 57, died, 62, 63, residence, 64, 65
49	officer, officers, police, fired, attempted, shot, times, began, involved, stop
50	white, died, 45, 48, 47, 53, 43, 38, 42, residence
51	residence, foul_play, deceased, circumstance, investigators, adult, firearm, evidence, signs, approximately
52	depression, medication, taking, history, anxiety, medications, prescribed, suffered, doctor, diagnosed
53	made, suicide, past, threats, suicidal, depressed, kill_himself, told, wanted, talked
54	argument, arguing, began, drinking, leave, threatened, started, left, argued, house
55	medical, history, diabetes, hypertension, disease, pressure, due, suffered, high_blood, significant
56	days, welfare_check, check, contact, requested, residence, deceased, called, unable,

	landlord
57	vehicle, car, driver, parked, truck, driving, drove, seat, road, side
58	deceased, white, residence, 30, scene, medical, history, 40, 31, suffered
59	wound, head, gunshot, handgun, pistol, weapon, semi_automatic, gun, firearm, exit
60	witness, bar, people, party, involved, witnesses, fight, group, altercation, street
61	scene, death, discovered, recovered, observed, responded, unknown, pronounced, related, manner
62	medical, emergency, service, scene, pronounced, responded, unresponsive, circumstanced, experiencing, unmentioned
63	note, suicide, notes, left, written, letter, life, addressed, family, stating
64	history, mental_health, disorder, depression, bipolar_disorder, treatment, diagnosed, schizophrenia, facility, bi_polar
65	emergency, hospital, transported, medical, room, service, pronounced, arrival, department, shortly_after
66	died, 22, 20, white, 24, 23, 21, incident, report, case
67	residence, white, 50, 51, 59, inside, scene, complainant, incident, discovered
68	information, incident, provided, noted, 25, 26, cousin, unknown, residence, died
69	prior, incident, day, months, weeks, month, week, days, approximately, released
70	medications, prescription, prescribed, clonazepam, depression, included, gabapentin, trazodone, citalopram, alprazolam
71	circumstances, prior, unknown, white, history, home, source, mental_health, treatment, clear
72	officers, related_party, hours, noted, medical, located, advised, examiner, stated, responded
73	shot, gun, wound, head, died, african_american, own_life, indicate_why, homocide, killed
74	death, manner, medical, examiner, 49, 41, coroner, ruled, dead, arrived
75	left, wound, back, death, entrance, brain, exit, bullet, upper, front
76	suicide, head, white, basement, bag, plastic_bag, note, death, discovered, secured
77	female, boyfriend, children, child, 39, relationship, exboyfriend, woman, custody, pregnant
78	gunshot, wound, head, self_inflicted, rifle, 22_caliber, heard, intra_oral, single, chin
79	missing, body, reported, area, park, wooded_area, located, woods, water, river
80	alcohol, drinking, fiance, beer, alcoholic, bottle, intoxicated, abuse, alcoholism, drank
81	law_enforcement, arrived, called, stated, called_911, call, advised, deceased, 911, laying
82	died, circumstances, result, manner, day, prior, death, responded, notes, services
83	friend, friends, house, told, night, day, staying, called, asked, time
84	mother, home, told, grandmother, day, uncle, mom, stepfather, aunt, earlier
85	police, called, arrived, told, call, responded, received, find, 911, department
86	prior, days, years, 52, 55, approximately, months, past, due, weeks

87	wound, chest, shotgun, self_inflicted, gunshot, head, weapon, 12_gauge, single, legs
88	bathroom, blood, floor, left, bathtub, wrist, cut, wrists, noted, shower
89	death, manner, toxicology, positive, blood, ethanol, autopsy, arrived, talked, showed
90	female, husband, white, home, exhusband, kill_herself, problems, history, estranged, attempted
91	scene, dead, pronounced, called, white, 27, 32, 29, 33, 34
92	suicide, note, left, exwife, depressed, committed, commit, stating, depression, contents
93	2, 1, 3, deceased, due, 4, residence, information, time, hispanic
94	jail, arrested, prison, charges, released, court, arrest, cell, case, probation
95	multiple, black, wounds, homicide, gunshot, suspect, street, suffering, motive, information
96	garage, suicide, car, inside, carbon_monoxide, poisoning, running, white, note, vehicle
97	daughter, heard, called, phone, told, called_911, house, home, arrived, upstairs
98	victim1, victim2, extra_victim2, victims, merged_into, victim3, female, shot, wounds, killed
99	injuries, head, blunt_force, trauma, jumped, struck, multiple, bridge, hit, train

Most representative terms are listed in order of highest to lowest probability of being generated by each LDA topic. Misspellings which were not caught in preprocessing are retained here.

The model with 100 topics has a coherence of 0.18 and topic diversity of 0.49. These metrics suggest that the LDA model captures a more limited and broad (i.e., less coherent and distinctive) set of topics compared to those picked up by the Discourse Atom Topic Model. The fact that these models pick up different kinds of topical structures is illustrated not only by metrics of coherence and diversity, but also by manually inspecting the topics. See all LDA topics in Table 7; this Table includes the topic number and the top 10 most representative words (by probability) for each LDA topic. In comparison with topics picked up by the Discourse Atom Topic Model, these topics tend towards more macro-level themes rather than the nuanced, focused topics identified by the Discourse Atom Topic Model; this is also indicated by the lower average topic coherence. For comparability, we also describe an LDA model with 225 topics (since our Discourse Atom Topic Model used 225 topics). For an LDA model with 225 topics, coherence is 0.18 and diversity is 0.41. We provide 15 randomly selected topics among these 225 topics in Table 8.

LDA topics are not immediately compatible with approaches to identify semantic dimensions in semantic space, e.g., finding a gender dimension, as we do in this paper. However, in datasets (like the NVDRS) that include structured variables (like victim gender) for each document, we *can* examine the prevalence of LDA topics by structured variables. Here, we do so by computing the mean proportion of each topic among female victims, and then doing so again among male victims. Then we divide these two distributions element-wise to identify the topics which are most distinct to women versus men. Note that this is the same procedure we used in the main text to construct the “Gender Prevalence Ratio” in Figure 6.

Using this approach, we find that the LDA topics that are most prevalent in female victims' narratives (versus those of male victims) are LDA topics 90, 77, 9, 70 and 98. Topics 90 and 77 appear to be about interpersonal violence, romantic relationships, and children (for the most representative words, see Table 7). Topics 9 and 70 are about prescription medications, and Topic 98 is about multiple victims. The LDA topics that are most distinctive to male victims' narratives (versus those of female victims) are topics 19, 12, 95, 18, and 78. Topics 19 and 12 are also about romantic relationships and family, but more focused on separation. Topic 95 is about homicides, and topics 18 and 78 are about gunshot wounds and gun actions. Importantly, this approach only captures the *prevalence* of LDA topics by gender. With DATM topics, we can also examine how these topics are gendered in semantic space, i.e., the gendered meaning in their typical language context.

Table 8. Random Sample of 15 Topics identified using LDA Topic Modeling with 225 Topics

Topic Number	**Top 10 Most Representative Terms**
38	room, hotel, motel, hotel_room, checked, motel_room, manager, staying, day, bed
44	witness, heard, witnesses, gunshots, stated, ground, street, area, scene, ran
87	white, died, 67, 68, residence, suicide, home, 75, grandson, 80,
94	police, white, park, death, report, 65, passerby, scene, state, deceased
97	shot, gun, head, wound, self_inflicted, shoot_himself, self_inflcited, tested_negative, himslef
107	stated, spoke, wanted, thought, knew, past, lot, aware, mentioned, hurt
111	job, lost, depressed, recently, due, losing, work, quit, loss, unemployed
123	argument, arguing, fiance, argued, night, gotten_into, drinking, left, leave, began
124	told, wanted, called, phone, talking, talked, talk, day, die, spoke
149	heroin, unresponsive, history, drug, cocaine, intoxication, drug_paraphernalia, fentanyl, syringe, morphine
152	incident, prior, day, 47, night, days, fatal, date, took_place, informed
163	3, 2, 4, 5, months, years, 6, hours, prior, approximately
166	details, unspecified, died, firearm, time, place, residence, body, age, location
185	hospital, transported, died, local, admitted, injuries, ambulance, expired, transferred, surgery
191	medical, emergency, service, scene, pronounced, notified, considered, clear_whether, moments_before, pronounces
Most representative terms are listed in order of highest to lowest probability of being generated by each LDA topic. Misspellings which were not caught in preprocessing are retained here.	

Semantic Dimensions Beyond Gender

As we emphasize in the main text, methods used to identify the biases or cultural connotations of words can be successfully extended to identify biases of topics extracted using DATM. In the main text, we illustrate this important extension of embedding methods using the case of gender. Here, we offer an additional application to illustrate this point: the extent to which topics are associated with descriptions of indoors or outdoors in the narratives. Indeed, this core approach to measure bias or cultural meaning in topics can be used for any strong, stable semantic contrast, including contrasts that may have theoretical motivation but are not (yet) cleanly represented in structured variables.

As with the construct of gender, we extract a dimension for indoors versus outdoors in the corpus. Specifically, we average the vectors for the words: indoors, inside, and indoor, and then subtract out the average of the vectors for the words: outdoors, outside, and outdoor. We examine the topics that load most highly onto the resulting dimension (i.e., have the highest or lowest cosine similarity). Topics with large negative cosine similarity are more distinct to language about outdoors (and not indoors), while topics with large positive cosine similarities are more distinct to language about indoors (and not outdoors). In our data, the most “outdoors” topic is one we had labeled “Specific outdoor locations” (Topic 120), followed by topics labeled “Rural outdoor areas” (Topic 219), and “Canvassed” (Topic 52). The most “indoors” topic is one labeled “Fumes” (Topic 14), followed by topics about “Tubes” (Topic 214), and “Gun Actions” (Topic 152). For context, fumes and tubes both reflect gas poisonings, which occur in closed (i.e., indoor) spaces.

Chapter 5. Conclusion

Across three papers, this dissertation contributes to methodological and theoretical work on meaning in cultural sociology, mathematical sociology, and computer science. The first paper bridges cultural sociology theory on meaning (namely, linguistic structuralism) with a burgeoning method to measure meaning in text data in social science (word embeddings). Word embedding methods are increasingly popular in sociology, but thus far are largely used as methods for empirical work. This paper highlights that an underrecognized value of word embedding methods is that they offer social scientists an opportunity to reimagine the study of meaning. It compares word embedding to a linguistic structuralist vision of meaning revealed new solutions to long standing theoretical critiques of linguistic structuralism itself. The future directions outlined in this paper further highlighted the many fruitful intersections between method and theory in the study of meaning. While this paper focused on the case of language, it also emphasized that takeaways from language can generalize to how meaning works (and can be measured) through other modalities as well.

Papers two and three focused on two cases where theoretical and methodological advances on meaning can aid empirical discovery (about gendered language and gendered patterns in violent death, respectively). Gendered language is critical to study because it reinforces and contributes to patterns of gender inequality through gendered meanings around occupations. Empirically, Paper two found that gendered meanings of words are linked to their suffixed form in a range of cultural texts. Theoretically, it also offered an account of how the relationship between gendered meaning and form in public culture might be learned and internalized in personal culture (Lizardo 2017). It then used word embedding methods to empirically model this process. More specifically, I use word embedding methods that account for word form (e.g., morphemes, suffixes, and affixes). For linguistic debates about gendered language, an important takeaway is that making language gender-neutral (or gender fair) is not just about changing which words (or words' forms) are used but changing *how* words are used. For cultural sociology, the empirical results in this paper challenge the longstanding notion that symbols' forms are arbitrarily related to their meanings.

Paper three is motivated by the concrete problem that analysts increasingly need to measure meaning in text at scales which are only possible with computational methods. Paper three develops a novel, robust way to identify latent topics in large-scale text data: Discourse Atom Topic Modeling. Using this new topic model, the paper then describes a large untapped source of information about violence in the U.S., which the CDC has collected since 2003. We identify a range of latent topics, some of which are not yet described in longstanding classification schemes for violent death. We further illustrate that word embedding and topic modeling approaches to analyze text data can now be combined in our model. For example, we describe several topics that are differentially used to describe women and men in narratives about violent death. While women and men are known to have concretely different patterns of violent deaths, this work did not distinguish gender differences in reporting versus gender differences in actual death instances. More broadly, like other large-scale administrative text data (e.g., clinical notes) this text data on violence remains greatly underutilized due to methodological barriers. The model introduced in paper three offers new solutions to overcome several of these longstanding barriers to measuring meaning in text data.

Finally, this dissertation and the future research directions outlined in each paper underscore the interdisciplinarity of questions about meaning. Each paper in this dissertation drew from a range of fields to theorize and provide new metrics for meaning — from cognitive science and linguistics to applied text analysis and theoretical work in machine-learning. The notion of semantic information is also fundamental not only to sociology, but psychology, linguistics, natural language processing, and information science. The lack of clarity around meaning is even considered a key barrier to developing artificial intelligence (Mitchell 2020). Across these fields, semantic information is conceptualized and measured in countless ways: from stereotypes and implicit bias to word co-occurrences, and points and lines in a high-dimensional semantic space. Future work on meaning can continue to benefit from bridging methods, empirical findings, and theory across these fields. Several, among many questions of interest to sociologists include: How is meaning represented in our minds and how might it be modeled in artificial minds? In what way are contextual cues used to process meaning? How does meaning travel between minds, or between public and personal culture? How does it manifest in external cultural objects, and in what ways does modality matter? How does meaning in personal culture correspond to and diverge from meaning in public culture? To what extent and how does meaning reflect material patterns? And finally, but perhaps most importantly for sociology, how does meaning guide action and contribute to inequality?

WORKS CITED

Aharon, Michal, Michael Elad, and Alfred Bruckstein. 2006. “K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation.” *IEEE Transactions on Signal Processing* 54(11):4311–22.

Akbik, Alan, Duncan Blythe, and Roland Vollgraf. 2018. “Contextual String Embeddings for Sequence Labeling.” Pp. 1638–49 in, *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Aletras, Nikolaos, and Mark Stevenson. 2013. “Evaluating Topic Coherence Using Distributional Semantics.” Pp. 13–22 in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)--Long Papers*.

Alexander, Jeffrey. 2003. *The Meanings of Social Life*. Oxford: Oxford University Press.

Alexander, Jeffrey, and Philip Smith. 1993. “The Discourse of American Civil Society: A New Proposal for Cultural Studies.” *Theory and Society* 22(2):151–207.

Antoniak, Maria, and David Mimno. 2018. “Evaluating the Stability of Embedding-Based Word Similarities.” *Transactions of the Association for Computational Linguistics* 6:107–19.

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. “A Practical Algorithm for Topic Modeling with Provable Guarantees.” Pp. 280–88 in *International Conference on Machine Learning*.

Arora, Sanjeev, Rong Ge, and Ankur Moitra. 2012. “Learning Topic Models--Going beyond SVD.” Pp. 1–10 in *2012 IEEE 53rd annual symposium on foundations of computer science*. IEEE.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. “A Latent Variable Model Approach to Pmi-Based Word Embeddings.” *Transactions of the Association for Computational Linguistics* 4:385–99.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Ma Tengyu, and Andrej Risteski. 2018. “Linear Algebraic Structure of Word Senses, with Applications to Polysemy.” *Transactions of the Association for Computational Linguistics* 6:483–95.

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2017. “A Simple but Tough-to-Beat Baseline for Sentence Embeddings.” Pp. 1–11 in *5th International Conference on Learning Representations, ICLR*.

Arora, Simran, Avner May, Jian Zhang, and Christopher Ré. 2020. “Contextual Embeddings: When Are They Worth It?” *ArXiv:2005.09117 [Cs]*.

Arronte Alvarez, Aitor, and Francisco Gómez-Martin. 2019. “Distributed Vector Representations of Folksong Motifs.” Pp. 325–32 in *Mathematics and Computation in Music*, edited by M. Montiel, F. Gomez-Martin, and O. A. Agustín-Aquino. Cham: Springer International Publishing.

Arseniev-Koehler, Alina, Susan D. Cochran, Vickie M. Mays, Kai-Wei Chang, and Jacob G. Foster. 2022. “Integrating Topic Modeling and Word Embedding to Characterize Violent Deaths.” *Proceedings of the National Academy of Sciences* 119(10):e2108801119. doi: 10.1073/pnas.2108801119.

Arseniev-Koehler, Alina, and Jacob G. Foster. 2020. “Machine Learning as a Model for Cultural Learning: Teaching an Algorithm What It Means to Be Fat.” *ArXiv* 2003(12133):1–45. doi: 10.31235/osf.io/c9yj3.

Arseniev-Koehler, Alina, Jacob Gates Foster, Vickie M. Mays, Kai-Wei Chang, and Susan D. Cochran. 2021. “Aggression, Escalation, and Other Latent Themes in Legal Intervention Deaths of Non-Hispanic Black and White Men: Results From the 2003–2017 National Violent Death Reporting System.” *American Journal of Public Health* e1–9. doi: 10.2105/AJPH.2021.306312.

Atagi, Natsuki, Nitya Sethuraman, and Linda B. Smith. 2009. “Conceptualizations of Gender in Language.” in *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 31.

Bail, Christopher. 2014. “The Cultural Environment: Measuring Culture with Big Data.” *Theory and Society* 43(3–4):465–82.

Bailey, April H., Marianne LaFrance, and John F. Dovidio. 2019. “Is Man the Measure of All Things? A Social Cognitive Account of Androcentrism.” *Personality and Social Psychology Review* 23(4):307–31. doi: 10.1177/1088868318782848.

Bakhtin, Mikhail M. 1981. “The Dialogic Imagination: Four Essays, Ed.” *Michael Holquist, Trans. Caryl Emerson and Michael Holquist (Austin: University of Texas Press, 1981)* 84(8):80–82.

Bamler, Robert, and Stephen Mandt. 2017. “Dynamic Word Embeddings.” Pp. 380–89 in *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Sydney, Australia.

Barber, C., D. Azrael, and D. Cohen. 2016. “Homicides by Police: Comparing Counts from the National Violent Death Reporting System, Vital Statistics, and Supplementary Homicide Reports.” *American Journal of Public Health* 106(5):922–27.

Baroni, Marco. 2016. “Grounding Distributional Semantics in the Visual World.” *Language and Linguistics Compass* 10(1):3–13.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewsk. 2014. “Don’t Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors.” Pp. 238–47 in *52nd Annual Meeting of the Association for Computational Linguistics*. Vol. 1.

Barsalou, Lawrence W. 1999. “Perceptual Symbol Systems.” *Behavioral and Brain Sciences* 22(4):577–660. doi: 10.1017/S0140525X99002149.

Barthes, Roland. 1961. “Toward a Psychosociology of Contemporary Food Consumption.” in *Food and culture*, edited by C. Counihan and P. V. Esterik. New York, NY: Routledge.

Barthes, Roland. 1977. *Elements of Semiology*. 34. [print]. New York, NY: Hill and Wang.

Basov, Nikita, Ronald Breiger, and Iina Hellsten. 2020. “Socio-Semantic and Other Dualities.” *Poetics* 78:101433.

Batton, Candice. 2004. “Gender Differences in Lethal Violence: Historical Trends in the Relationship between Homicide and Suicide Rates, 1960–2000.” *Justice Quarterly* 21(3):423–61.

Bem, Sandra L., and Daryl J. Bem. 1973. “Does Sex-biased Job Advertising ‘Aid and Abet’ Sex Discrimination?” *Journal of Applied Social Psychology* 3(1):6–18.

Best, Rachel Kahn, and Alina Arseniev-Koehler. 2022. *Stigma’s Uneven Decline*. *preprint*. SocArXiv. doi: 10.31235/osf.io/7nm9x.

Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, and Aleksandr Nisnevich. 2020. “Experience Grounds Language.” Pp. 8718–35 in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Blasi, Damián E., Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. 2016. “Sound–Meaning Association Biases Evidenced across Thousands of Languages.” *Proceedings of the National Academy of Sciences* 113(39):10818–23.

Blei, David. 2012. “Probabilistic Topic Models.” *Communications of the ACM* 55(4):77–84.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. “Language (Technology) Is Power: A Critical Survey of ‘Bias’ in NLP.” Pp. 5454–76 in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. “Enriching Word Vectors with Subword Information.” *Transactions of the Association for Computational Linguistics* 5:135–46.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings.” Pp. 4349–57 in *Advances in neural information processing systems*. Barcelona, Spain.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan

Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. "On the Opportunities and Risks of Foundation Models." *ArXiv:2108.07258 [Cs]*.

Borgatti, Stephen P., Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. 2009. "Network Analysis in the Social Science." *Science* 323(5916):892–95.

Borghi, Anna, Ferdinand Binkofski, Cristiano Castelfranchi, Felice Cimatti, Claudia Scorolli, and Luca Tummolini. 2017. "The Challenge of Abstract Concepts." *Psychological Bulletin* 143:263.

Boroditsky, Lera, Lauren A. Schmidt, and Webb Phillips. 2003. "Sex, Syntax, and Semantics." Pp. 61–79 in *Language in mind: Advances in the study of language and thought*, edited by D. Gentner and S. Goldin-Meadow. Cambridge, Massachusetts: The MFT Press.

Bourdieu, Pierre. 1984. *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press.

Boutyline, Andrei, Alina Arseniev-Koehler, and Devin Cornell. 2020. "School, Studying, and Smarts: Gender Stereotypes and Education Across 80 Years of American Print Media, 1930-2009." *SocArXiv*. doi: 10.31235/osf.io/bukdg.

Boutyline, Andrei, Devin Cornell, and Alina Arseniev-Koehler. 2021. "All Roads Lead to Polenta: Cultural Attractors at the Junction of Public and Personal Culture." *Sociological Forum* 36(S1):1419–45. doi: 10.1111/socf.12760.

Boyd-Graber, Jordan, David Mimno, and David Newman. 2014. "Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements." P. 225255 in *Handbook of mixed membership models and their applications*.

Brekhus, Wayne. 1998. "A Sociology of the Unmarked: Redirecting Our Focus." *Sociological Theory* 16(1):34–51.

Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. 2014. "Multimodal Distributional Semantics." *Journal of Artificial Intelligence Research* 49(2014):1–47.

Bryson, Joanna. 2008. "Embodiment versus Memetics." *Mind & Society* 7(1):77–94.

Calanca, Federica, Luiza Sayfullina, Lara Minkus, Claudia Wagner, and Eric Malmi. 2019. "Responsible Team Players Wanted: An Analysis of Soft Skill Requirements in Job Advertisements." *EPJ Data Science* 8(1):13.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-like Biases." *Science* 356(6334):183–86.

Caliskan, Aylin, and Molly Lewis. 2022. "Social Biases in Word Embeddings and Their Relation to Human Cognition." Pp. 447–63 in *Handbook of Language Analysis in Psychology*. New York: Guilford Publications.

Callanan, Valerie, and Mark Davis. 2012. "Gender Differences in Suicide Methods." *Social Psychiatry and Psychiatric Epidemiology* 47(6):857–69.

Centers for Disease Control and Prevention. 2019. *National Violent Death Reporting System (NVDRS)*. National Center for Injury Prevention and Control, Division of Violence Prevention.

Cerulo, Karen A. 1995. *Identity Designs: The Sights and Sounds of a Nation*. Rutgers University Press.

Cerulo, Karen A. 2018. "Scents and Sensibility: Olfaction, Sense-Making, and Meaning Attribution." *American Sociological Review* 83(2):361–89. doi: 10.1177/0003122418759679.

Cerulo, Karen A. 2019. "Embodied Cognition." *The Oxford Handbook of Cognitive Sociology* 81.

Chakrabarti, Parijat, and Margaret Frye. 2017. "A Mixed-Methods Framework for Analyzing Text Data: Integrating Computational Techniques with Qualitative Methods in Demography." *Demographic Research* 37:1351–82.

Charlesworth, Tessa ES, Victor Yang, Thomas C. Mann, Benedek Kurdi, and Mahzarin R. Banaji. 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):218–40.

Chen, Dawn, Joshua C. Peterson, and Thomas L. Griffiths. 2017. "Evaluating Vector-Space Models of Analogy." *ArXiv* 1705(04416).

Chuan, Ching-Hua, Kat Agres, and Dorien Herremans. 2020. "From Context to Concept: Exploring Semantic Relationships in Music with Word2vec." *Neural Computing and Applications* 32(4):1023–36. doi: 10.1007/s00521-018-3923-1.

Craib, I. 1992. "The Word as a Logical Pattern: An Introduction to Structuralism." Pp. 131–48 in *Modern Social Theory*. St Martin's Press.

Das, Rajarashi, Manzil Zaheer, and Chris Dyer. 2015. "Gaussian Lda for Topic Models with Word Embeddings." Pp. 795–804 in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Vol. 1.

Davis, Charles P., and Eiling Yee. 2021. "Building Semantic Memory from Embodied and Distributional Language Experience." *WIREs Cognitive Science* 12(5). doi: 10.1002/wcs.1555.

Dehghani, Morteza, Reihane Boghrati, Kingson Man, Joe Hoover, Sarah I. Gimbel, Ashish Vaswani, Jason D. Zevin, Mary Helen Immordino-Yang, Andrew S. Gordon, Antonio Damasio, and Jonas T. Kaplan. 2017. "Decoding the Neural Representation of Story Meanings across Languages." *Human Brain Mapping* 38(12):6096–6106.

Deutscher, Guy. 2010. *Through the Language Glass: Why the World Looks Different in Other Languages*. Metropolitan books.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Pp. 4171–86 in *NAACL-HLT*.

Dieng, Adji B., Francisco JR Ruiz, and David M. Blei. 2020. "Topic Modeling in Embedding Spaces." *Transactions of the Association for Computational Linguistics* 8:439–53.

DiMaggio, Paul. 1997. "Culture and Cognition." *Annual Review of Sociology* 23(1):263–87.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41:570–606.

Dingemanse, Mark, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen, and Padraic Monaghan. 2015. "Arbitrariness, Iconicity, and Systematicity in Language." *Trends in Cognitive Sciences* 19(10):603–15. doi: 10.1016/j.tics.2015.07.013.

Dosse, François. 1997. *History of Structuralism*. Vol. 1. Minneapolis, Minn: University of Minnesota Press.

Douglas, Mary. 2003. *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. Routledge.

Eagly, Alice H., and Valerie J. Steffen. 1984. "Gender Stereotypes Stem from the Distribution of Women and Men into Social Roles." *Journal of Personality and Social Psychology* 46(4):735–54. doi: 10.1037/0022-3514.46.4.735.

Eisenstein, Jacob, Amr Ahmed, and Eric P. Xing. 2011. "Sparse Additive Generative Models of Text." Pp. 1041–48 in *Proceedings of the 28th international conference on machine learning (ICML-11)*.

Emirbayer, Mustafa. 2004. "The Alexander School of Cultural Sociology." *Thesis Eleven* 79(1):5–15. doi: 10.1177/0725513604046951.

Ertl, Allison, Kameron Sheats, Emiko Petrovsky, Carter Betz, Keming Yuan, and Katherine Fowler. 2019. "Surveillance for Violent Deaths - National Violent Death Reporting System, 32 States, 2016." *MMWR Surveillance Summaries* 68(9):1.

Ethayarajh, Kawin. 2018. "Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline." Pp. 91–100 in *Proceedings of The Third Workshop on Representation Learning for NLP*.

Ethayarajh, Kawin. 2019. “How Contextual Are Contextualized Word Representations?” in *IJCNLP*.

Ethayarajh, Kawin, David Duvenaud, and Graeme Hirst. 2019. “Understanding Undesirable Word Embedding Associations.” *ArXiv Preprint ArXiv:1908.06361*.

Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. “Problems With Evaluation of Word Embeddings Using Word Similarity Tasks.” Pp. 30–35 in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. “Placing Search in Context: The Concept Revisited.” *ACM Transactions on Information Systems* 20(1):116–31.

Firth, John. 1957. “A Synopsis of Linguistic Theory, 1930-1955.” Pp. 1–32 in *Studies in Linguistic Analysis*. Oxford: Philological Society.

Formanowicz, Magdalena, Sylwia Bedynska, Aleksandra Cisłak, Friederike Braun, and Sabine Sczesny. 2013. “Side Effects of Gender-fair Language: How Feminine Job Titles Influence the Evaluation of Female Applicants.” *European Journal of Social Psychology* 43(1):62–71.

Foster, Jacob G. 2018. “Culture and Computation: Steps to a Probably Approximately Correct Theory of Culture.” *Poetics* 68:144–54.

Fox, James Alan, and Emma E. Fridel. 2017. “Gender Differences in Patterns and Trends in US Homicide, 1976–2015.” *Violence and Gender* 4(2):37–43.

Franco, Karlien, Dirk Geeraerts, Dirk Speelman, and Roeland Van Hout. 2019. “Concept Characteristics and Variation in Lexical Diversity in Two Dutch Dialect Areas.” *Cognitive Linguistics* 30(1):205–42.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. “Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes.” *Proceedings of the National Academy of Sciences* 115(16):E3635–44.

Geeraerts, Dirk. 2010. *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Geeraerts, Dirk, Stefan Grondelaers, and Peter Bakema. 2012. *The Structure of Lexical Variation: Meaning, Naming, and Context*. Vol. 5. Walter de Gruyter.

Geeraerts, Dirk, and Dirk Speelman. 2010. “Heterodox Concept Features and Onomasiological Heterogeneity in Dialects.” Pp. 21–40 in *Advances in cognitive sociolinguistics*. De Gruyter Mouton.

Geertz, Clifford. 1973. “Thick Description: Toward an Interpretive Theory of Culture.” Pp. 3–30 in *The interpretation of cultures: selected essay*. New York: Basic Books.

Ghaziani, Delia, Amin Baldassarri. 2011. “Cultural Anchors and the Organization of Differences.” *American Sociological Review* 76(2):179–206.

Giddens, Anthony. 1979. *Central Problems in Social Theory: Action, Structure and Contradiction in Social Analysis*. Berkeley, CA: University of California Press.

Gladkova, Anna, and Aleksandr Drozd. 2016. "Intrinsic Evaluations of Word Embeddings: What Can We Do Better?" Pp. 36–42 in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany: Association for Computational Linguistics.

Goh, Gabriel, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. "Multimodal Neurons in Artificial Neural Networks." *Distill* 6(3):10.23915/distill.00030. doi: 10.23915/distill.00030.

Goldstick, Jason E., Patrick M. Carter, and Rebecca M. Cunningham. 2021. "Current Epidemiological Trends in Firearm Mortality in the United States." *JAMA Psychiatry* 78(3):241–42.

Grand, Gabriel, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2018. "Semantic Projection: Recovering Human Knowledge of Multiple, Distinct Object Features from Word Embeddings." *ArXiv Preprint ArXiv:1802.01241*.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan LK Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74(6):1464.

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

Grover, Aditya, and Jure Leskovec. 2016. "Node2vec: Scalable Feature Learning for Networks." Pp. 855–64 in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM.

Günther, Fritz, Luca Rinaldi, and Marco Marelli. 2019. "Vector-Space Models of Semantic Representation from a Cognitive Perspective: A Discussion of Common Misconceptions." *Perspectives on Psychological Science* 14(6):1006–33.

Haber, Jaren R. 2021. "Sorting Schools: A Computational Analysis of Charter School Identities and Stratification." *Sociology of Education* 94(1):43–64. doi: 10.1177/0038040720953218.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." Pp. 1489–1501 in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Han, Lushan, Abhay L. Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. "UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems." Pp. 44–52 in, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Atlanta, Georgia, USA: Association for Computational Linguistics.

Hanlon, Thomas J., Catherine Barber, Deborah Azrael, and Matthew Miller. 2019. "Type of Firearm Used in Suicides: Findings from 13 States in the National Violent Death Reporting System, 2005–2015." *Journal of Adolescent Health* 65(3):366–70.

Harris, Zellig. 1954. "Distributional Structure." *Word* 10(2–3):146–62.

Haspelmath, Martin. 2006. "Against Markedness (and What to Replace It With)." *Journal of Linguistics* 42(1):25–70.

He, Guimei. 2010. "An Analysis of Sexism in English." *Journal of Language Teaching and Research* 1(3):332–35. doi: 10.4304/jltr.1.3.332-335.

Hellinger, M., and H. Bußmann. 2003. *Gender Across Languages: The Linguistic Representation of Women and Men*. Amsterdam: John Benjamins Publishing Company.

Hemenway, David, Catherine Barber, and Matthew Miller. 2010. "Unintentional Firearm Deaths: A Comparison of Other-Inflicted and Self-Inflicted Shootings." *Accident Analysis & Prevention* 42(4):1184–88.

Hepburn, Lisa, Matthew Miller, Deborah Azrael, and David Hemenway. 2007. "The US Gun Stock: Results from the 2004 National Firearms Survey." *Injury Prevention* 13(1):15–19.

Herz, Marcus, and Thomas Johansson. 2015. "The Normativity of the Concept of Heteronormativity." *Journal of Homosexuality* 62(8):1009–20. doi: 10.1080/00918369.2015.1021631.

Hinton, Geoffrey E., and Russ R. Salakhutdinov. 2009. "Replicated Softmax: An Undirected Topic Model." *Advances in Neural Information Processing Systems* 22.

Hirschberg, Julia, and Christopher D. Manning. 2015. "Advances in Natural Language Processing." *Science* 349(6245):261–66.

Hockett, Charles F. 1960. "The Origin of Speech." *Scientific American* 203(3):88–97.

Hollis, Geoff. 2017. "Estimating the Average Need of Semantic Knowledge from Distributional Semantic Models." *Memory & Cognition* 45(8):1350–70. doi: 10.3758/s13421-017-0732-1.

Holmes, Janet. 2001. "A Corpus-Based View of Gender in New Zealand English." *Gender across Languages: The Linguistic Representation of Women and Men* 1:115–36.

Holmes, Janet, and Robert Sigley. 2002. "What'sa Word like Girl Doing in a Place like This? Occupational Labels, Sexist Usages and Corpus Research." Pp. 247–63 in *New frontiers of corpus research*. Brill.

Hu, Tianran, Ruihua Song, Maya Abtahian, Philip Ding, Xing Xie, and Jiebo Luo. 2017. "A World of Difference: Divergent Word Interpretations among People." in *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11.

Ignatow, G. 2009. "Culture and Embodied Cognition: Moral Discourses in Internet Support Groups for Overeaters." *Social Forces* 88(2):643–69. doi: 10.1353/sof.0.0262.

Ignatow, Gabe. 2016. “Theoretical Foundations for Digital Text Analysis.” *Journal for the Theory of Social Behaviour* 46(1):104–20.

Ignatow, Gabriel. 2007. “Theories of Embodied Knowledge: New Directions for Cultural and Cognitive Sociology?” *Journal for the Theory of Social Behaviour* 37(2):115–35. doi: 10.1111/j.1468-5914.2007.00328.x.

Joas, H., and W. Knobl. 2009. “Structuralism and Poststructuralism.” Pp. 339–70 in *Social Theory: Twenty Introductory Lectures*. Cambridge University Press.

Jones, Frank L., and Philip Smith. 2001. “Diversity and Commonality in National Identities: An Exploratory Analysis of Cross-National Patterns.” *Journal of Sociology* 37(1):45–63. doi: 10.1177/144078301128756193.

Jones, Jason, Amin Ruhul Mohammad, Jessica Kim, and Steven Skiena. 2020. “Stereotypical Gender Associations in Language Have Decreased Over Time.” *Sociological Science* 7:1–35.

Joseph, Kenneth, and Jonathan Morgan. 2020. *When Do Word Embeddings Accurately Reflect Surveys on Our Beliefs About People?*

Kelly, Michael H. 1992. “Using Sound to Solve Syntactic Problems: The Role of Phonology in Grammatical Category Assignments.” *Psychological Review* 99(2):349.

Kim, Yoon, Yi-I. Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. “Temporal Analysis of Language through Neural Language Models.” Pp. 61–65 in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics.

Köhler, Wolfgang. 1929. *Gestalt Psychology*. New York: Liveright Publishing Corporation.

Kozlowski, Austin, Matt Taddy, and James A. Evans. 2019. “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings.” *American Sociological Review* 84(5):905–49.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. “Statistically Significant Detection of Linguistic Change.” Pp. 625–35 in *Proceedings of the 24th International Conference on World Wide Web*. Florence Italy: International World Wide Web Conferences Steering Committee.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. “Diachronic Word Embeddings and Semantic Shifts: A Survey.” *ArXiv:1806.03537 [Cs]*.

Labov, William. 1972. *Sociolinguistic Patterns*. University of Pennsylvania press.

Lakoff, George, and Mark Johnson. 1980. “Conceptual Metaphor in Everyday Language.” *The Journal of Philosophy* 77(8):453–86.

Lakoff, George, and Mark Johnson. 2008. *Metaphors We Live By*. University of Chicago press.

Landauer, Thomas, and Susan Dumais. 1997. “A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge.” *Psychological Review* (104):211–40.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. “An Introduction to Latent Semantic Analysis.” *Discourse Processes* 25(2–3):259–84.

Larsen, Anders Boesen Lindbo, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2015. “Autoencoding beyond Pixels Using a Learned Similarity Metric.” *International Conference on Machine Learning* 48:1558–66.

Lembo, Alessandra, and John Levi Martin. 2021. “The Structure of Cultural Experience.” *Poetics* 101562. doi: 10.1016/j.poetic.2021.101562.

Leminen, Alina, Eva Smolka, Jon A. Dunabeitia, and Christos Pliatsikas. 2018. “Morphological Processing in the Brain: The Good (Inflection), the Bad (Derivation) and the Ugly (Compounding).” *Cortex* 116:4–44.

Lenci, Alessandro. 2008. “Distributional Semantics in Linguistic and Cognitive Research.” *Italian Journal of Linguistics* 20(1):1–31.

Lenci, Alessandro. 2018. “Distributional Models of Word Meaning.” *Annual Review of Linguistics* (4):151–71.

Leschziner, Vanina, and Gordon Brett. 2021. “Symbol Systems and Social Structures.” Pp. 559–82 in *Handbook of Classical Sociological Theory*. Cham: Springer.

Lévi-Strauss, Claude. 1963. *Structural Anthropology*. New York, London: Basic Books.

Levy, Omer, and Yoav Goldberg. 2014. “Neural Word Embedding as Implicit Matrix Factorization.” *Advances in Neural Information Processing Systems* 2177–85.

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. “Improving Distributional Similarity with Lessons Learned from Word Embeddings. Transactions of the Association for Computational Linguistics.” *Transactions of the Association for Computational Linguistics* 3:211–25.

Lewis, Molly, Matt Cooper Borkenhagen, Ellen Converse, Gary Lupyan, and Mark S. Seidenberg. 2022. “What Might Books Be Teaching Young Children About Gender?” *Psychological Science* 33(1):33–47. doi: 10.1177/09567976211024643.

Lewis, Molly, and Gary Lupyan. 2020. “Gender Stereotypes Are Reflected in the Distributional Structure of 25 Languages.” *Nature Human Behaviour* 1–8.

Li, Lucy, and Jon Gauthier. 2017. “Are Distributional Representations Ready for the Real World? Evaluating Word Vectors for Grounded Perceptual Meaning.” *ArXiv Preprint* 1705(11168).

Liben, Lynn S., Rebecca S. Bigler, and Holleen R. Krogh. 2002. “Language at Work: Children’s Gendered Interpretations of Occupational Titles.” *Child Development* 73(3):810–28.

Lindqvist, Anna, Emma Aurora Renström, and Marie Gustafsson Sendén. 2018. "Reducing a Male Bias in Language? Establishing the Efficiency of Three Different Gender-Fair Language Strategies." *Sex Roles* 1–9.

Linzhuo, Li, Wu Lingfei, and James Evans. 2020. "Social Centralization and Semantic Collapse: Hyperbolic Embeddings of Networks and Text." *Poetics* 78:101428. doi: 10.1016/j.poetic.2019.101428.

Lizardo, Omar. 2016. "Cultural Symbols and Cultural Power." *Qualitative Sociology* 39(2):199–204. doi: 10.1007/s11133-016-9329-4.

Lizardo, Omar. 2017. "Improving Cultural Analysis: Considering Personal Culture in Its Declarative and Nondeclarative Modes." *American Sociological Review* 82(1):88–115.

Lynott, Dermot, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. "The Lancaster Sensorimotor Norms: Multidimensional Measures of Perceptual and Action Strength for 40,000 English Words." *Behavior Research Methods* 52(3):1271–91. doi: 10.3758/s13428-019-01316-z.

Mandera, Pawel, Emmanuel Keuleers, and Marc Brysbaert. 2017. "Explaining Human Performance in Psycholinguistic Tasks with Models of Semantic Similarity Based on Prediction and Counting: A Review and Empirical Validation." *Journal of Memory and Language* 92:57–78.

Martin, John Levi. 2010. "Life's a Beach but You're an Ant, and Other Unwelcome News for the Sociology of Culture." *Poetics* 38(2):229–44.

Martin, Paul, and Pam Papadelos. 2017. "Who Stands for the Norm? The Place of Metonymy in Androcentric Language." *Social Semiotics* 27(1):39–58. doi: 10.1080/10350330.2016.1145371.

Martin-Caughey, Ananda. 2021. "What's in an Occupation? Investigating Within-Occupation Variation and Gender Segregation Using Job Titles and Task Descriptions." *American Sociological Review* 86(5):960–99. doi: 10.1177/00031224211042053.

Maryanski, Alexandra, and Jonathan Turner. 1992. "The Offspring of Functionalism: French and British Structuralism." *Sociological Theory* 9(1):106–15.

McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit."

McConnell, Allen R., and Russell H. Fazio. 1996. "Women as Men and People: Effects of Gender-Marked Language." *Personality and Social Psychology Bulletin* 22(10):1004–13. doi: 10.1177/01461672962210003.

Mecit, Alican, L. J. Shrum, and Tina M. Lowrey. 2022. "COVID-19 Is Feminine: Grammatical Gender Influences Danger Perceptions and Precautionary Behavioral Intentions by Activating Gender Stereotypes." *Journal of Consumer Psychology* 32(2):316–25. doi: 10.1002/jcpy.1257.

Merchant, Amil, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. "What Happens To BERT Embeddings During Fine-Tuning?" Pp. 33–44 in *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics.

Merton, Robert K. 1948. “The Bearing of Empirical Research on Sociological Theory.” *American Sociological Review* 13(5):505–15.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *ArXiv [Cs.CL]* 1301.3781.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. “Advances in Pre-Training Distributed Word Representations.” *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” *Advances in Neural Information Processing Systems* 3111–19.

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. 2013. “Linguistic Regularities in Continuous Space Word Representations.” Pp. 746–51 in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*.

Mohr, John. 1998. “Measuring Meaning Structures.” *Annual Review of Sociology* 24(1):345–70.

Mohr, John, Chris Bail, Margaret Frye, Jennifer Lena, Omar Lizardo, Terence McDonnell, Ann Mische, Iddo Tavory, and Frederick Wherry. 2020. *Measuring Culture*. New York: Columbia University Press.

Mohr, John, and Amin Ghaziani. 2014. “Problems and Prospects of Measurement in the Study of Culture.” *Theory and Society* 43(3–4):226–46.

Monaghan, Padraic, Morten H. Christiansen, and Nick Chater. 2007. “The Phonological-Distributional Coherence Hypothesis: Cross-Linguistic Evidence in Language Acquisition.” *Cognitive Psychology* 55(4):259–305.

Morris, Wayne. 2013. “Transforming Able-Bodied Normativity: The Wounded Christ and Human Vulnerability.” *Irish Theological Quarterly* 78(3):231–43. doi: 10.1177/0021140013484428.

Moseley, Rachel, Francesca Carota, Olaf Hauk, Bettina Mohr, and Friedemann Pulvermüller. 2012. “A Role for the Motor System in Binding Abstract Emotional Meaning.” *Cerebral Cortex* 22(7):1634–47. doi: 10.1093/cercor/bhr238.

Mu, Jiaqi, Suma Bhat, and Pramod Viswanath. 2017. “All-but-the-Top: Simple and Effective Postprocessing for Word Representations.” *ArXiv Preprint ArXiv:1702.01417*.

Murphy, Sherry L., Jiaquan Xu, Kenneth D. Kochanek, and Elizabeth Arias. 2018. “Mortality in the United States, 2017.”

Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. "Efficient Non-Parametric Estimation of Multiple Embeddings per Word in Vector Space." *ArXiv Preprint* 1504(06654).

Nelson, Laura K. 2021. "Leveraging the Alignment between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century US South." *Poetics* 101539.

Nickel, Maximillian, and Kiela Kiela. 2017. "Poincaré Embeddings for Learning Hierarchical Representations." Pp. 6338–47 in *Advances in neural information processing systems*.

Nosek, Brian A., Mahzarin R. Banaji, and Anthony G. Greenwald. 2002. "Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Web Site." *Group Dynamics: Theory, Research, and Practice* 6(1):101.

O'callaghan, Derek, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. "An Analysis of the Coherence of Descriptors in Topic Modeling." *Expert Systems with Applications* 42(13):5645–57.

Pati, Yagyensh Chandra, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. 1993. "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition." Pp. 40–44 in *Proceedings of 27th Asilomar conference on signals, systems and computers*. IEEE.

Paulozzi, Leonard J., J. Mercy, Lorraine Frazier, and J. Lee Annest. 2004. "CDC's National Violent Death Reporting System: Background and Methodology." *Injury Prevention* 10(1):47–52.

Peirsman, Yves, Dirk Geeraerts, and Dirk Speelman. 2010. "The Automatic Identification of Lexical Variation between Language Varieties." *Natural Language Engineering* 16(4):469–91.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global Vectors for Word Representation." Pp. 1532–43 in *Conference on empirical methods in natural language processing (EMNLP)*.

Pérez, Efrén O., and Margit Tavits. 2019. "Language Influences Public Attitudes toward Gender Equality." *The Journal of Politics* 81(1):81–93.

Perniss, Pamela, Robin L. Thompson, and Gabriella Vigliocco. 2010. "Iconicity as a General Property of Language: Evidence from Spoken and Signed Languages." *Frontiers in Psychology* 1:227.

Peters, Matthew, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. "Semi-Supervised Sequence Tagging with Bidirectional Language Models." Pp. 1756–65 in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics.

Peters, Matthew E., Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. "Dissecting Contextual Word Embeddings: Architecture and Representation." *ArXiv:1808.08949 [Cs]*.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. “Deep Contextualized Word Representations.” Pp. 2227–37 in.

Petterson, James, Wray Buntine, Shravan Narayanamurthy, Tibério Caetano, and Alex Smola. 2010. “Word Features for Latent Dirichlet Allocation.” in *Advances in Neural Information Processing Systems*. Vol. 23, edited by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. Curran Associates, Inc.

Phillips, Webb, and Lera Boroditsky. 2003. “Can Quirks of Grammar Affect the Way You Think? Grammatical Gender and Object Concepts.” *Proceedings of the Annual Meeting of the Cognitive Science Society* 12:928–33.

Pulvermüller, Friedemann. 2013. “How Neurons Make Meaning: Brain Mechanisms for Embodied and Abstract-Symbolic Semantics.” *Trends in Cognitive Sciences* 17(9):458–70. doi: 10.1016/j.tics.2013.06.004.

Quiroga, R. Quian, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. 2005. “Invariant Visual Representation by Single Neurons in the Human Brain.” *Nature* 435(7045):1102–7.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. “Learning Transferable Visual Models From Natural Language Supervision.” *ArXiv:2103.00020 [Cs]*.

Ramachandran, Vilayanur S., and Edward M. Hubbard. 2001. “Synaesthesia–a Window into Perception, Thought and Language.” *Journal of Consciousness Studies* 8(12):3–34.

Řehůřek, Radim, and Petr Sojka. 2010. “Software Framework for Topic Modelling with Large Corpora.” Pp. 45–50 in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA.

Reynolds, David, Alan Garnham, and Jane Oakhill. 2006. “Evidence of Immediate Activation of Gender Information from a Social Role Name.” *Quarterly Journal of Experimental Psychology* 59(5):886–903. doi: 10.1080/02724980543000088.

Ridgeway, Cecilia L. 2011. *Framed by Gender: How Gender Inequality Persists in the Modern World*. Oxford: Oxford University Press.

Ridgeway, Cecilia L., and Shelley J. Correll. 2004. “Unpacking the Gender System: A Theoretical Perspective on Gender Beliefs and Social Relations.” *Gender & Society* 18(4):510–31.

Robinson, Justyna A. 2010. “Awesome Insights into Semantic Variation.” Pp. 85–110 in *Advances in cognitive sociolinguistics*. De Gruyter Mouton.

Rockett, Ian RH, Eric D. Caine, Hilary S. Connery, Gail D’Onofrio, David J. Gunnell, Ted R. Miller, Kurt B. Nolte, Mark S. Kaplan, Nestor D. Kapusta, Christa L. Lilly, and others. 2018. “Discerning Suicide in Drug Intoxication Deaths: Paucity and Primacy of Suicide Notes and Psychiatric History.” *PLoS One* 13(1):e0190200.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures." Pp. 399–408 in *Proceedings of the eighth ACM international conference on Web search and data mining*.

Rong, Xin. 2014. "Word2Vec Parameter Learning Explained." *ArXiv* 1411–2738.

Rosenberg, Mark L., Lucy E. Davidson, Jack C. Smith, Alan L. Berman, Herb Buzbee, George Gantner, George A. Gay, Barbara Moore-Lewis, Don H. Mills, Don Murray, and others. 1988. "Operational Criteria for the Determination of Suicide." *Journal of Forensic Science* 33(6):1445–56.

Rosenfeld, Alex, and Katrin Erk. 2018. "Deep Neural Models of Semantic Shift." Pp. 474–84 in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics.

Roy, Deb. 2005. "Grounding Words in Perception and Action: Computational Insights." *Trends in Cognitive Sciences* 9:389–96.

Rozado, David. 2019. "Using Word Embeddings to Analyze How Universities Conceptualize 'Diversity' in Their Online Institutional Presence." *Society* 56(3):256–66. doi: 10.1007/s12115-019-00362-9.

Rozado, David, and Musa al-Gharbi. 2021. "Using Word Embeddings to Probe Sentiment Associations of Politically Loaded Terms in News and Opinion Articles from News Media Outlets." *Journal of Computational Social Science*. doi: 10.1007/s42001-021-00130-y.

Rubenstein, Ron, Michael Zibulevsky, and Michael Elad. 2008. "Efficient Implementation of the K-SVD Algorithm Using Batch Orthogonal Matching Pursuit." *CS Technion Report, Computer Science Department, Technion*.

Sahlgren, Magnus. 2008. "The Distributional Hypothesis." *Italian Journal of Disability Studies* 20:33–53.

Samuel, Steven, Geoff Cole, and Madeline J. Eacott. 2019. "Grammatical Gender and Linguistic Relativity: A Systematic Review." *Psychonomic Bulletin & Review* 26(6):1767–86. doi: 10.3758/s13423-019-01652-3.

Saussure, Ferdinand de. 1916. "The Linguistic Sign." Pp. 24–46 in *Semiotics: An Introductory Anthology*, edited by R. Innis. Bloomington: Indiana University Press.

Saussure, Ferdinand de. 1983. *Course in General Linguistics*. London: Duckworth.

Schofield, Alexandra, M\aans Magnusson, and D. Mimno. 2017. "Understanding Text Pre-Processing for Latent Dirichlet Allocation." Pp. 432–36 in *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*. Vol. 2.

Schofield, Alexandra, Mans Magnusson, and David Mimno. 2017. "Pulling Out the Stops: Rethinking Stopword Removal for Topic Models." *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* 2:423–36.

Sczesny, Sabine, Magda Formanowicz, and Franziska Moser. 2016. "Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination?" *Frontiers in Psychology* 7:25.

Sera, Maria D., Christian AH Berge, and Javier del Castillo Pintado. 1994. "Grammatical and Conceptual Forces in the Attribution of Gender by English and Spanish Speakers." *Cognitive Development* 9(3):261–92.

Sewell, W. H. 2005. "The Concept(s) of Culture." Pp. 152–74 in *The Logics of History*. University of Chicago Press.

Sewell, William H. 1992. "A Theory of Structure: Duality, Agency, and Transformation." *American Journal of Sociology* 98(1):1–29.

Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. "Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview." Pp. 5248–64 in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Smith, Linda, and Michael Gasser. 2005. "The Development of Embodied Cognition: Six Lessons from Babies." *Artificial Life* 11(1–2):13–29.

Sperber, Dan. 1985. "Anthropology and Psychology: Towards an Epidemiology of Representations." *Man* 73–89.

Stites, Mallory C., Kara D. Federmeier, and Kiel Christianson. 2016. "Do Morphemes Matter When Reading Compound Words with Transposed Letters? Evidence from Eye-Tracking and Event-Related Potentials." *Language, Cognition and Neuroscience* 31(10):1299–1319.

Stoltz, Dustin. 2019. "Becoming a Dominant Misinterpreted Source: The Case of Ferdinand de Saussure in Cultural Sociology." *Journal of Classical Sociology* 0(0):1–22.

Stoltz, Dustin S., and Marshall A. Taylor. 2021. "Cultural Cartography with Word Embeddings." *Poetics* 101567.

Stoltz, Dustin, and Marshall Taylor. 2019. "Concept Mover's Distance: Measuring Concept Engagement via Word Embeddings in Texts." *Journal of Computational Social Science* 2(2):293–313.

Stoltz, Dustin, and Marshall Taylor. 2020. "Cultural Cartography with Word Embeddings." *ArXiv Preprint* 2007.04508:1–70.

Stone, Deborah M., Kristin M. Holland, Brad Bartholow, Joseph E. Logan, Wendy LiKamWa McIntosh, Aimee Trudeau, and Ian RH Rockett. 2017. "Deciphering Suicide and Other Manners of Death Associated with Drug Intoxication: A Centers for Disease Control and Prevention Consultation Meeting Summary." *American Journal of Public Health* 107(8):1233–39.

Stout, Jane G., and Nilanjana Dasgupta. 2011. "When He Doesn't Mean You: Gender-Exclusive Language as Ostracism." *Personality and Social Psychology Bulletin* 37(6):757–69.

Strauss, Claudia, and Naomi Quinn. 1997. *A Cognitive Theory of Cultural Meaning*. Vol. 9. Cambridge, United Kingdom: Cambridge University Press.

Swidler, Ann. 1986. “Culture in Action: Symbols and Strategies.” *American Sociological Review* 51:273–86.

Swidler, Ann. 2013. *Talk of Love: How Culture Matters*. University of Chicago Press.

Talbert, Ryan D. 2017. “Culture and the Confederate Flag: Attitudes toward a Divisive Symbol.” *Sociology Compass* 11(2). doi: 10.1111/soc4.12454.

Tavory, Iddo, and Ann Swidler. 2009. “Condom Semiotics: Meaning and Condom Use in Rural Malawi.” *American Sociological Review* 74(2):171–89. doi: 10.1177/000312240907400201.

Taylor, Marshall A., and Dustin S. Stoltz. 2021. “Integrating Semantic Directions with Concept Mover’s Distance to Measure Binary Concept Engagement.” *Journal of Computational Social Science* 4(1):231–42.

Taylor, Marshall, and Dustin Stoltz. 2020. “Concept Class Analysis: A Method for Identifying Cultural Schemas in Texts.” *Sociological Science* 7:544–69.

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. “What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations.” *ArXiv:1905.06316 [Cs]*.

Vervecken, Dries, Pascal Gygax, Ute Gabriel, Matthias Guillod, and Bettina Hannover. 2016. “Warm-Hearted Businessmen, Competitive Housewives? Effects of Gender-Fair Language on Adolescents’ Perceptions of Occupations.” *Frontiers in Psychology* 6:1437.

Vervecken, Dries, Bettina Hannover, and Ilka Wolter. 2013. “Changing (S) Expectations: How Gender Fair Job Descriptions Impact Children’s Perceptions and Interest Regarding Traditionally Male Occupations.” *Journal of Vocational Behavior* 82(3):208–20.

Vijayakumar, Ashwin K., Ramakrishna Vedantam, and Devi Parikh. 2017. “Sound-Word2Vec: Learning Word Representations Grounded in Sounds.” *ArXiv:1703.01720 [Cs]*.

Wang, Bin, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C. C. Jay Kuo. 2019. “Evaluating Word Embedding Models: Methods and Experimental Results.” *APSIPA Transactions on Signal and Information Processing* 8.

Wang, Yuxuan, Yutai Hou, Wanxiang Che, and Ting Liu. 2020. “From Static to Dynamic Word Representations: A Survey.” *International Journal of Machine Learning and Cybernetics* 11(7):1611–30. doi: 10.1007/s13042-020-01069-8.

Whorf, Benjamin. 1956. *Language, Thought and Reality: Selected Writing of Benjamín Lee Whorf.* Cambridge, MA: MIT Press.

Xie, Pengtao, Diyi Yang, and Eric Xing. 2015. "Incorporating Word Correlation Knowledge into Topic Modeling." Pp. 725–34 in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics.

Xu, Huimin, Zhang Zhang, Lingfei Wu, and Cheng-Jun Wang. 2019. "The Cinderella Complex: Word Embeddings Reveal Gender Stereotypes in Movies and Books" edited by I. Safro. *PLOS ONE* 14(11):e0225385. doi: 10.1371/journal.pone.0225385.

Yakin, Halina Sendera Mohd, and Andreas Totu. 2014. "The Semiotic Perspectives of Peirce and Saussure: A Brief Comparative Study." *Procedia-Social and Behavioral Sciences* 155:4–8.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. "Dynamic Word Embeddings for Evolving Semantic Discovery." Pp. 673–81 in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. Marina Del Rey CA USA: ACM.

Zerubavel, Eviatar. 2018. *Taken for Granted: The Remarkable Power of the Unremarkable*. Princeton, NJ: Princeton University Press.

Zhao, He, Lan Du, and Wray Buntine. 2017. "A Word Embeddings Informed Focused Topic Model." Pp. 423–38 in *Asian Conference on Machine Learning*.