# The Geometry of Culture: Analyzing Meaning through Word Embeddings

Austin C. Kozlowski[1]
Matt Taddy[2,3]
James A. Evans[1]

[1] University of Chicago, Department of Sociology
[2] University of Chicago, Booth School of Business
[3] Amazon

We demonstrate the utility of a new methodological tool, neural-network word embedding models, for large-scale text analysis, revealing how these models produce richer insights into cultural associations and categories than possible with prior methods. Word embeddings represent semantic relations between words as geometric relationships between vectors in a high-dimensional space, operationalizing a relational model of meaning consistent with contemporary theories of identity and culture. We show that dimensions induced by word differences (e.g., *man – woman*, *rich – poor*, *black – white*, *liberal – conservative*) in these vector spaces closely correspond to dimensions of cultural meaning, and the projection of words onto these dimensions reflects widely shared cultural connotations when compared to surveyed responses and labeled historical data. We pilot a method for testing the stability of these associations, then demonstrate applications of word embeddings for macro-cultural investigation with a longitudinal analysis of the coevolution of gender and class associations in the United States over the 20th century and comparative analysis of historic distinctions between markers of gender and class in the U.S. and Britain. We argue that the success of these high-dimensional models motivates a move towards "high-dimensional theorizing" of meanings, identities and cultural processes.

## INTRODUCTION

A vast amount of information about what people do, know, think, and feel lies preserved in digitized text and an increasing proportion of social life now occurs natively in this medium. Available sources of digitized text are wide ranging, including collective activity on the web, social media, and instant messages as well as online transactions, medical records, and digitized letters, pamphlets, articles, and books (Evans and Aceves 2016; Grimmer and Stewart 2013). This growing supply of text has elicited demand for natural language processing and machine-learning tools to filter, search, and translate text into valuable data. The analysis of large digitized corpora has already proven fruitful in a range of social scientific endeavors including analysis of discourse surrounding political elections and social movements, the accumulation of knowledge in the production of science, and communication and collaboration within organizations (Christopher Bail 2012; Evans and Aceves 2016; Foster, Rzhetsky, and Evans 2015; Grimmer 2009; Goldberg et al. 2016).

Although text analysis has long been a cornerstone for the study of culture, the impact of "big data" on the sociology of culture remains modest (Bail 2014). A fundamental challenge for the computational analysis of text is to simultaneously leverage the richness and complexity inherent in large corpora while producing a representation simple enough to be intuitively understandable, analytically useful and theoretically relevant. Moreover, turning text into data (Grimmer and Stewart 2013) requires credible methods for (1) evaluating the statistical significance of observed patterns and (2) disciplining the space of interpretations to avoid the tendency to creatively confirm expectations (Nickerson 1998). While past research has made strides towards overcoming these challenges in the study of culture, critics continue to argue that existing methods fail to capture the nuances of text that can be gleaned from interpretive text analysis (Biernacki 2012).

In this paper, we demonstrate the utility of a new computational approach – neural-network word embedding models – for the sociological analysis of culture. We show that word embedding models are able to capture more complex semantic relations than past modes of computational text analysis and can prove a powerful tool in the study of cultural categories and associations. Word embeddings are high-dimensional vector-space models[1] of text in which each unique word in the corpus is represented as a vector in a shared vector space (Mikolov, Yih, and Zweig 2013; Pennington, Socher, and Manning 2014). Methods similar to word embeddings, such as Latent Semantic Analysis (LSA) or Indexing (LSI), have existed in various forms since the 1970s (Dumais 2004). Recent breakthroughs in auto-encoding neural networks and advances in computational power have enabled a new class of word embedding models that incorporate relevant information about word contexts from highly local windows of surrounding words rather than an entire surrounding document. As a result, these new word embedding models distil an encyclopedic breadth of subtle and complex cultural associations from large collections of text by training the model with local word associations a human might learn through ambient exposure to the same collection of language.

In word embedding models, words are assigned a position in a vector space based on the context that word shares with other words in the corpus. Words that share many contexts are positioned near one another, while words that inhabit very different contexts locate farther apart. Previous work with word embedding models in computational linguistics and natural language processing has shown that words frequently sharing linguistic contexts, and thus located nearby in the vector space, tend to share similar meanings. However, semantic information is encoded not only in the clustering of words with similar meanings. In this paper we present evidence that the very dimensions of these vector space models closely correspond to meaningful "cultural

[1] Word embedding models are sometimes considered and referred to as "low dimension" techniques relative to the number of words used in text (e.g., 20,000) because they reduce this *very* high dimensional word space. Nevertheless, considered from the perspective of one, two or three dimensional models common in the analysis of culture, these spaces are much more complex, and reproduce much more accurate total associations, as shown below.

dimensions" such as race, class, and gender. We show that the positioning of word vectors along culturally salient dimensions within the vector space captures how concepts are related to one another within cultural categories. For example, projecting occupation names on a "gender dimension," we find that traditionally feminine occupations such as "nurse" and "nanny" are positioned at one end the dimension and traditionally masculine occupations such as "engineer" and "lawyer" are positioned at the opposite end. This occurs because with each local context that "nurse" shares with feminine words like "she," "her," and "woman," it is nudged towards the feminine pole of the gender dimension, while each time "engineer" shares a context with terms like "his," "him," and "man," it is nudged toward the masculine pole.

We emphasize that word embedding models not only grant insight into the semantic structure of a given cultural system, but can also be productively used to analyze cultural difference and change. By comparing word embedding models trained on texts produced in different cultural contexts, it is possible to directly examine differences in the in the meanings of terms and categories between social groups. Similarly, comparing word embedding models trained on texts from sequential time periods make it possible to investigate how cultural associations shift historically.

In the following analyses, we utilize word embedding models trained on contemporary texts, texts stretching back over a century, and texts from distinct cultures to demonstrate the broad potential of word embedding models for cultural analysis. The following analyses proceed in four steps. First, we compare semantic relations derived from word embedding models to results from survey data measuring cultural associations in order to demonstrate the ecological validity of word embedding models in capturing widespread associations with race, class, and gender. Second, we analyze word embedding models trained on texts published in the United States spanning the entire 20th century to investigate macro-historic cultural trends, discovering slow but persistent changes in the interrelationship of the categories of gender and class. Third, we conduct a cross-national analysis, comparing text from the United States and Great Britain from the turn of the 20th century to identify subtle differences in the markers of class and gender between these two social contexts. The findings presented here demonstrate the broad utility of word embedding models for cultural analysis and motivate further development and application of these models to a diverse array of questions in sociology of culture.

## FORMAL TEXT ANALYSIS IN THE STUDY OF CULTURE

Language has long held a central position in the study of culture. Cultural scholars from sociology, anthropology, and socio-linguistics have commonly understood a group's language to be a reflection of its cultural system (Whorf 1956; Levi-strauss 1963), and thus text has served as a key source of data for scholars investigating cultural categories and meaning structures. Historically, analysis of text in sociology has been dominated by qualitative, interpretive approaches, the two most common being interpretivist close-reading and systematic qualitative coding. Interpretivist text analysis, in which the researcher draws insights from a holistic deep reading of text, has done much to advance sociological understandings of culture, but suffers from clear limitations in reproducibility (Ricoeur 1981). The method of qualitative coding, in which the researcher selects a number of themes and systematically tracks their deployment in text (Glaser and Strauss 1967) can be more reproducible than a singular close reading, but still suffers from low inter-coder reliability when themes are complex or subtle. Because these