THE UNIVERSITY OF CHICAGO

LINGUISTIC RELATIVITY, COLLECTIVE COGNITION, AND TEAM PERFORMANCE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF SOCIOLOGY

BY

PEDRO ACEVES

CHICAGO, ILLINOIS

JUNE 2018

Abstract

Linguistic relativity is the principle that the structure of the language that one speaks influences their cognition. This dissertation extends this principle into sociological territory, arguing that the structure of the language that a group speaks influences how its members interact. In the first chapter, I outline the new arsenal of computational tools that have been developed to work with the vast new stores of digital text data that have become available, and I highlight how these tools and data have been used by social scientists to glean insights about the social world. In the second chapter I put these tools to work on the development of a new language structure attribute, information density, which is the average amount of information contained within the words of a language. I show that there is significant variation in the rate of information density across languages, and that this rate is systematically related to rates of conceptual density and speech information transfer across languages. In the third chapter, I investigate whether the language information density rate is associated with the performance of expeditions to the Himalayas. I find that language information density is associated with greater expedition performance and with faster performance, especially within low-hierarchy teams where the benefits of communication would be most likely to be observed. This dissertation aims to advance a sociology of language wherein variation in language structure is the analytical input that shapes processes of social interaction, collective cognition, and group performance. It points to the importance of considering language structure as a significant force shaping social life.

TABLE OF CONTENTS

List of Tables

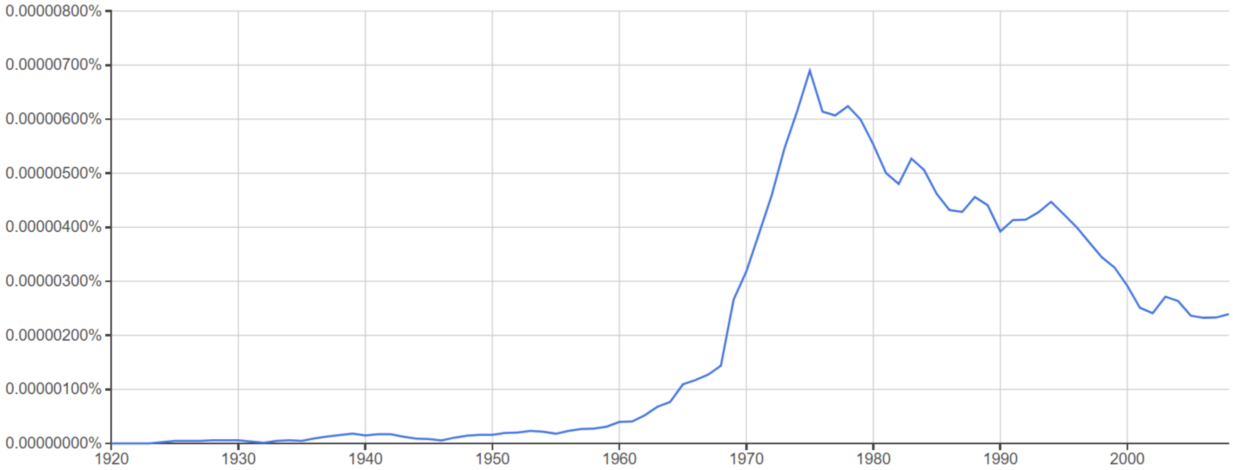
List of Figures

Introduction

Language is a fundamental aspect of social life. Without language, human behavior as we know it, from the simplest familial interactions to the grandest of scientific, economic, and military collaborations, would not be possible. As Searle has stated, “you can imagine a society that has language but has no government, property, marriage, or money. But you cannot imagine a society that has a government, property, marriage, and money but no language” (Searle 2010, p. 109). Given that language is such a central component of social reality, it is surprising that sociology has largely shied away from thinking about the interrelationship between language *qua* language and social life.

What explains the near absence of language from sociological theory? This gap may have arisen for several reasons. First, there was a general lack of attention paid to language by sociology’s founding fathers. As Coulmas has noted, Durkheim, Weber, and Parsons “took little interest in language as a social fundamental” (1998, p. 5). These scholars lived in monoglot societies, and to them language was omnipresent and invariant, and must have seemed of no significance in differentiating social behavior (Coulmas 1998; Fishman 1968). Second, the overarching concern with large-scale social structure and quantitative analysis made such research unlikely during the first decades of modern sociology (Fishman 1968). It is only when interest grew during the 1960’s and 1970’s in topics such as developing nations, small group dynamics, and network concomitants of unity and diversity that sociology of language started to emerge. Yet by the 1980’s and 1990’s interest again waned (See figure 0.1). During these years, the field of sociolinguistics, whose scholars were primarily housed in linguistics departments, continued to thrive. The primary interest of both the initial wave of the sociology of language and of sociolinguistics was to understand how the social factors of speech communities influenced their use of language. Third, the lack of research at the intersection of sociology and language might also be attributed to the inherent complexity of analytically understanding language. This difficulty has two sources. On the one hand, language is a complex, adaptive system and analytical clarity on it as an object of inquiry has grown in parallel with the emergence of sociology proper. It thus might have been too much to expect contributions at the intersection of these fields during their parallel emergence. On the other hand, the subfields within linguistics that have emerged since the 1960's have become ever more specialized. This confluence has made brokering the space between language and sociology difficult terrain to navigate. Finally, lack of research was likely hampered by the difficulty and expense of collecting and coding language data, especially for sociological research.

Figure 0.1: Appearance frequency of the terms "sociology of language" in the Goolge N-Gram Viewer



Within the past ten years, however, tools for the analysis of language as well as vast amounts of textual data for analysis have become prevalent. More of the social world lives within electronic text than ever before, from collective activity on the web, social media, and instant messaging to online transactions, government intelligence, and digitized libraries. This supply of text has elicited demand for natural language processing and machine learning tools to filter, search, and translate text into valuable data. Chapter 1 surveys some of the most exciting computational approaches to text analysis, highlighting both supervised methods that extend old theories to new data and unsupervised techniques that discover hidden regularities worth theorizing. It then reviews recent research that uses these tools to develop social insight by exploring (a) collective attention and reasoning through *content* from communication; (b) social relationships through the *process* of communication; and (c) social states, roles, and moves identified through heterogeneous *signals* within communication.

These new tools and data are highly suggestive of the broad space of possibility for brokering the intersection of language and social life. Over the last forty years the field of sociolinguistics has brokered this intersection by tracing a causal arrow from social structure and social facts broadly conceived to language structure and language use. Its emphasis has been on understanding the social determinants of language use by distinct speech communities (for reviews, see Bayley et al. 2013; Coulmas 1998; Holmes 2008). The goal of this dissertation is to broker the intersection between language and social life by tracing the arrow from language structure and language use to social action. This I hope will become the proper domain of the sociology of language.

In chapter 2, I ask whether lexicalized conceptual information is universally organized across human groups, or whether the nature of conceptual information varies across languages? Much work has investigated the encoding of conceptual information varies within locally circumscribed conceptual domains. Yet, little is known about the global organization of conceptual information across entire languages. Recent advances in digital data availability and machine learning algorithms now allow us to ask this question. Across a global sample of 986 languages representing 93 language families, I show that languages vary in how they organize conceptual information, displaying wide differences in the average amount of information that they encode within their words. I then show that the rate of language information density is positively associated with the density of the conceptual space within which concepts semantically relate to each other and negatively associated with the speech information transfer rate. These findings point to important differences within a central domain of human cognition and culture.

In chapter 3, I put the language information density measure to theoretical use in the context of team performance. Team performance research has established how variation in team structure or composition shapes performance outcomes. Both research perspectives posit that the flow of information among group members and between groups and their broader environment constitute primary mechanisms underlying collective performance. Language—the principal technology of information transfer—might be expected to play a critical role in group dynamics and performance. Nevertheless, virtually no research attention has examined the role that language structure plays in shaping group processes and outcomes. In this chapter, I examine whether differences in language information density lead to distinct performance outcomes. I argue that higher information density facilitates group movement through conceptual space as groups converse, especially during tasks requiring effective search, creativity, problem-solving, and decision-making. Tracing the outcomes of 1,626 monolingual expeditions to the Himalayas from 1907 to 2015, I show that speaking languages with higher information density is associated with summiting a larger proportion of team members, and conditional on reaching the summit, doing so more quickly. I extend insights from this study to explore how and why other contexts of group and organizational activity are likely to be influenced by language information density.

In chapter 4, I conclude by describing two experimental studies I have designed and am in the process of implementing that are intended to more clearly understand how it is that differences in language information density shape the nature of social interaction and collective cognition.

1 Text for Social Theory

A vast expanse of information about what people do, know, think, and feel lies embedded in text[1]. Textual traces range from the world's life on the web, social media, instant messaging and online commerce to automatically transcribed YouTube videos, medical records, digitized libraries and government intelligence. The rise of literacy, and more recently computers, scanners, the Internet and cellphones, have conditioned an exploding supply and demand for textual information. This provides sociologists access to a greater variety of texts that reach deeper into the contemporary social world than ever before. Simultaneously, massive semi-automated archival projects (e.g., GoogleBooks) are making vast caches of historical text digitally available for analysis.

This unfolding universe of digital text has generated a call for new information representations, natural language processing (NLP) and information retrieval (IR) and extraction (IE) tools that can filter, search, summarize, classify and extract information from text. Moreover, datasets representing not only the increased prevalence of text, but also audio, visual and heterogeneous sensor data (e.g., clickstreams on the web, "likes" on Facebook, movements via cellphone) have supported the rapid growth of a new engineering paradigm, machine learning (ML). An offspring of statistics and artificial intelligence, ML devotes itself to learning from data, predicting and extending human perceptive accuracy and understanding. Many general and text-specific ML techniques have now proven powerful for translating text and related communicative traces into sociologically valuable data (Grimmer & Stewart 2013).

[1] This chapter was previously published (Evans & Aceves 2016).

In this chapter, we briefly review the history of content and text analysis in sociology and the social sciences. Text is sometimes layers removed from the "social games" that sociologists seek to illuminate.[2] Computational approaches are sometimes less subtle and deep than the reading of a skillful analyst, who interprets text in context. Nevertheless, we show that recent advances in natural language processing and machine learning are being used to enhance qualitative analysis in two ways. First, supervised ML prediction tools can "learn" and reliably extend many sociologically interesting textual classifications to massive text samples far beyond human capacity to read, curate and code. Second, unsupervised ML approaches can "discover" unnoticed regularities in these massive samples of text that may merit sociological consideration and theorization.

Next, we review some of the most exciting computational approaches to the large-scale analysis of text for the production of sociologically relevant data. These include techniques from NLP, IR, IE, and ML that exploit language structure and context to extract meaning. Recent developments in ML, like the rise of "deep learning" or multi-layer neural networks, can change the technical machinery underlying these tools and improve their accuracy (Manning 2015). As a result, we focus on the persistent language tools and tasks (e.g., disambiguation, parsing) whose

[2] We do not use "Social games" to imply that human nature is primarily playful (Huizinga 1971) or that human action reflects exclusively rational, ends-oriented competition as in game theory (Myerson 2013; Neumann & Morgenstern 1944). Rather, like Bourdieu's field theory in which social agents play high stakes games of status on an established field (2013, p. 247), we intimate that social games comprise a board, pieces, conventional rules, established moves, and widely but not universally shared objectives that include both playing and winning. Wittgenstein's "language game" incorporates several of these aspects, but in a game restricted to the conveyance meaning through partially shared representations (2010, § 7). Social life, then, is composed of overlapping games (Long 1958), each with their own rules, in which players bring interests, dispositions, and strategies to their moves. Information about these games can be gleaned through the textual evidence they leave behind.

designs have proven valuable for the production of sociologically relevant data, despite changes in implementation.

The bulk of the chapter will focus on reviewing recent research that uses these approaches to develop social insight. This research groups itself into work that explores (1) collective attention and reasoning through examining the *content* of communication; (2) social relationships through analyzing the *process* of communication; and (3) social states, roles and moves identified through heterogeneous *signals* within communication. This work demonstrates large-scale analyses of not only human communication, but also the social and cultural worlds that produced it, and which become visible through it.

Content Analysis and Computation

What are the limits of text analysis? How and how well do text and other communicated content trace the social world that produced it? Here we take the "social world" to be comprised of individuals engaged in interaction, situated on a landscape characterized by (1) "social" structures like class boundaries, kin networks, formal organizations and ephemeral friendships; (2) "cultural" systems of shared symbols including the human communication protocols of language, gesture, and fashion; and (3) "material" or apparent external resources and constraints such as capital flows and the built environment. Life in this social world constitutes the "social game" in which individuals are engaged. For example, lawyers and judges play the legal game, students and teachers play the education game, politicians play the diplomacy game while generals play the war game. Actors also participate in a myriad of less specialized games including the parenting game, the courtship game, and the job market game (Long 1958).[3]

Individuals possess interests and drives that partially derive from the positions they assume within the social world, and which condition an unfolding stream of social action and interaction. These actions and interactions produce other- and self-communicated content, sometimes in the form of text, but also audio, video and even image recordings that can be translated with more or less fidelity into text. Consider transcripts of the tapes tracing deliberations among President Kennedy and his advisors surrounding the botched Bay of Pigs invasion and the Cuban Missile Crisis (Gibson 2012), or the Nixon White House tapes that detailed strategic discussion of the Watergate break-in and cover-up. It is not surprising that the games of Presidents have long been recorded, but the ubiquity of online communication, automated speech-to-text translation, and mobile sensors have made these traces available for a much wider range of social games and players than ever before.

This increase in available text has the potential to increase its relevance for many areas of sociological scholarship. Texts and communicative traces reveal more about some social games than others. A personal journal entry may uncover the state of the writer, while instant messaging banter reveals the intensity with which conversationalists initiate and maintain contact. While enabling certain views into the underlying social game, genres of text restrict others, as the modern research article, which obscures scientists' personal sensations and experiences by fixing on the referential world of experiment, observation and shared significance (Rodriguez-Esteban & Rzhetsky 2008; Shapin 1994, chapter 4).

[3] We use "social games" rather than "social interaction" or "social behavior", because the concept is more inclusive, comprising characteristics of the social players and environment, in addition to their communicative engagement.

In this way, inferences about the social, cultural and material landscape on which social games are played depend on the coupling between text and the underlying game. If details traced in text constitute substantial moves in the game, like “flirts” in an online dating website, then text may constitute a representative sample or even the complete population of relevant social moves. The more tightly coupled a text is to social moves in the game of interest, the stronger the inferences that can be made. The more loosely coupled—consider genres of memoir and hagiography—the stronger the assumptions required and the less relevant data possessed. The increase in textual data of all types allows us to reliably analyze games of greater structural and temporal complexity. Inferring differences, change over time or variation along some other dimension within social games or worlds requires more text than inferring stable, universal patterns. The deluge of historical and contemporary text digitally available today opens the possibility of inferring even more elaborate structures and patterns within social games, such as cycles, spatial arrays, complex hierarchies, and transitive orders (Kemp & Tenenbaum 2008).

Systematic text analysis entered sociology during the second meeting of the German Sociological Society in 1910 when Max Weber proposed a large-scale analysis of the German press to identify the influence of the news “in making modern man”, and trace temporal shifts in values (Hardt 2001, p. 136). This research program, interrupted by World War I, resurfaced in the quantitative analyses of mass media, including newspapers (Willey 1926; Woodward 1934), television and radio (Berelson & Lazarsfeld 1948). Microsociologists also began to analyze the content of group interactions (Bales 1950) and the structure of conversations (Sacks 1995 [1964]).

Sociologists engaging in content or conversation analysis often begin by qualitatively and reliably coding text or other media according to theoretically meaningful categories. They then quantitatively or interpretively analyze coded features, often in concert with raw textual elements (e.g., frequent or distinguishing words) and metadata (e.g., authorship, audience) to identify semantic or stylistic patterns. Consider Rosigno and Hodson's analysis of worker resistance through the coding of shop floor ethnographies (2004), or Stivers et al's analysis of transcribed interaction data across cultures to identify universal patterns in turn-taking (2009).

From the 1960s, computers have been used to assist sociologists in what a Rand Corporation paper termed *Automatic Content Analysis* (Hays 1960), beginning Philip Stone and Robert Bales' General Enquirer System, which mapped text to content dictionaries that tallied disambiguated words associated with power, sentiment and other categories tracing concepts relevant to theories in sociology, political science and psychology (Stone, Dunphy, Smith and Ogilvie 1966), much as systems created from subject-ranked terms, such as Tausczik and Pennebaker's LIWC (Linguistic Inquiry and Word Count) do today (2001; 2010). Later work by sociologist Kathleen Carley analyzed symbols in networks of associations to identify cultural patterns (Carley 1994) or what Sally Sedelow has called "society's collective associative memory" (1989).

In the last decade, however, statistical NLP has become dramatically more accurate and powerful at recovering linguistic structures and semantic associations recognized by both linguists and ordinary language speakers. Moreover, general machine learning models and algorithms have become much stronger in their ability to predict a range of outcomes, including expert annotations and underlying qualities of context through unstructured and semi-structured text data. Computational approaches can now make inferences substantially stronger with ML methods acting as extensions of our cognitive capacity—as cognitive prosthetics.

Computation can augment our fine perception of patterns in language and their links to the social world beneath. In past sociological work, a researcher might code passages of text relating to some underlying concept (e.g., feminism, democracy) or process (persuasion, consensus), but often without recognizing or articulating the details of language associated with those codes. They might not be able to construct a protocol enabling a naïve researcher or computer to reproduce them independently. ML approaches, coupled with a sufficiently rich set of textual features can be trained on human codes to extend them with improved fidelity. This can allow researchers to automatically analyze many, many more documents than would be possible through traditional reading. For example, manually coding topics from 40 million scientific abstracts could take a thousand researcher-years, but auto-coding them with a trained model might require only a few computer-days.

Moreover, ML techniques can be used to detect and predict qualities of the author, audience and social world from textual details imperceptible to human researchers. Machine memory augments human analytical limits by holding a massive array of language features simultaneously in mind so we can associate them in reliable "constant comparison" (Glaser 1965). Finally, ML tools can discover novel patterns in text data, based on similarity, structural association, or predictive power. These may merit interpretive scrutiny, labeling and incorporation into social theory. In summary, although computation cannot mimic the prior experience, vision and unexpected associations of a gifted analyst, it can augment their reliability and provide new data—regularities, associations and structures built from much larger text samples—which sociology can mine to deepen and expand our inferences about the social games and worlds underlying communication.

Data mining has acquired a bad reputation in the social sciences. Many see it as synonymous with the practice of algorithmically sifting through data for associations, then falsely reporting them as if confirmations of theoretically inspired, single-test hypotheses. Unreported and statistically unaccountable data mining leads to the over fitting of statistical models to data and fragile findings that neither replicate nor generalize. This, in turn, undermines confidence in published social science research (Freese 2007; Ioannidis & Doucouliagos 2013; Simmons et al. 2011a), just as it has in other fields like genetics, biomedicine (Ioannidis 2005) and even machine learning itself (Pentland 2012b). Unreported data mining was especially problematic in a social science era more heavily reliant on sparse, expensive data like in-person surveys and experiments. During such a time, not only published inferences, but potential data reuses were compromised by mining data's structure before social theories could be blind-tested against it. With the greater volume and variety of text data available today, produced both passively through the natural flow of digital communication and controlled online experiments, these concerns should be ameliorated: statistically accountable data mining can be used to legitimately *discover hypotheses* on some data and then *confirm those hypotheses* on other data. As such, with the contemporary explosion of text and the socially relevant data being mined from it, we expect to see a renaissance of discovery about human communication and the myriad social structures and processes reflected in it.

Data in Text

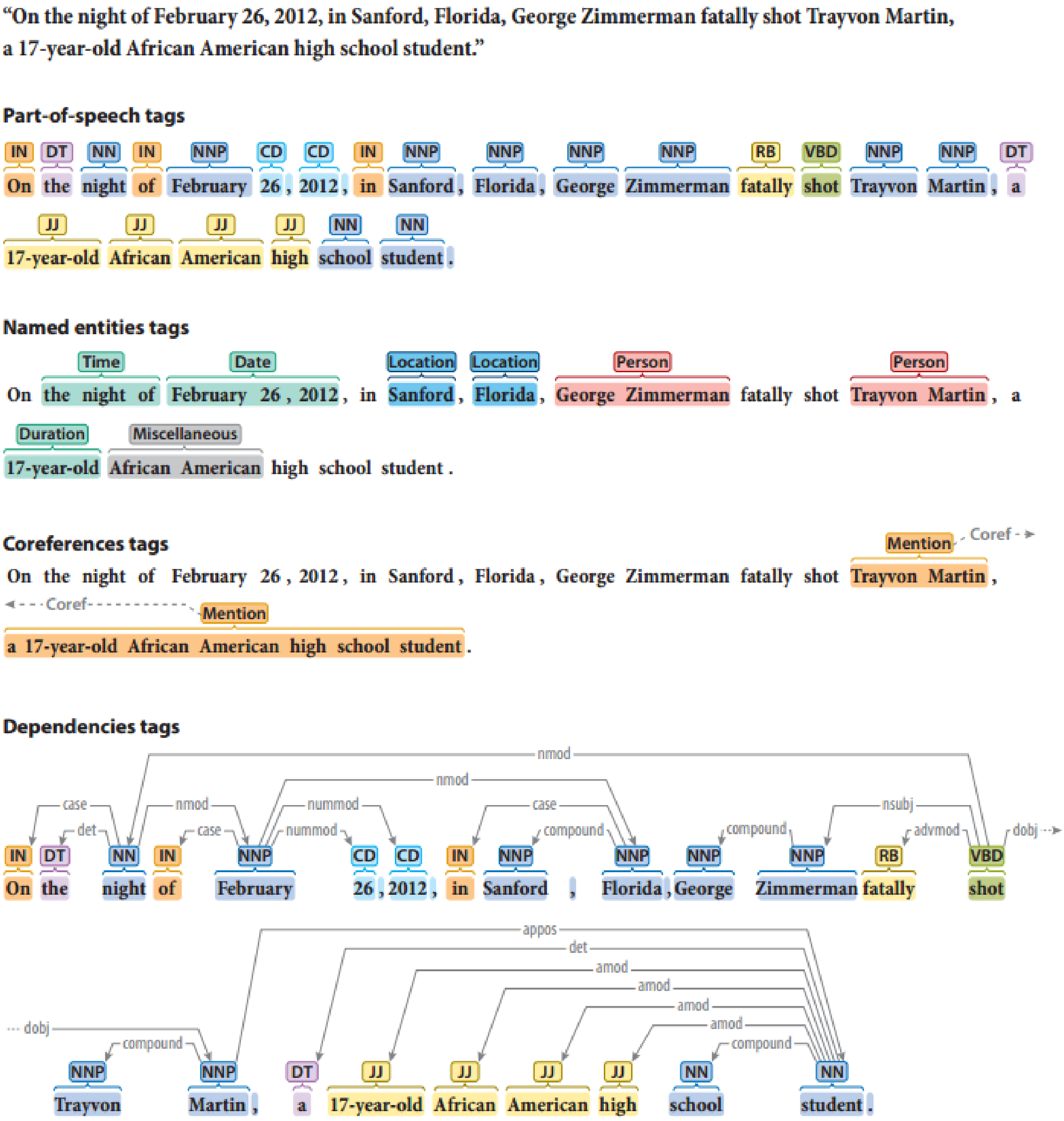
Text analysis attends to a range of language features, each of which condition modes of analysis with techniques from NLP, IR and IE. We cannot provide more than a dense, cursory treatment of these in this chapter; interested readers could consult the following book-length references for more complete (and relaxed) explication (Clark et al. 2010; Jurafsky & Martin 2000; Manning et al. 2008; Manning & Schütze 1999). In Figure 1, we quote a news source about the killing of Trayvon Martin alongside some prominent language features widely considered in text analysis, each extracted automatically from Stanford CoreNLP (Manning et al. 2014).

Most common is the *lexicon*—words in the vocabulary under consideration. Many studies use some function of word frequency to make inferences about meaning and focus. For example, analysts have used topically-curated word lists to identify states like ideology and emotion (Stone et al. 1966; Tausczik & Pennebaker 2010; Whissell 1989). Alternatively, the field of stylistics relies on the usage pattern of function words like articles and prepositions that carry no independent semantic information, in order to predict authorship through distinctive statistical signatures (Mosteller & Wallace 1964) or trace power dynamics by tracking mimicry (Danescu-Niculescu-Mizil et al. 2012). Words can also be associated by the role they play within a sentence through part-of-speech tagging.

When the lexicon is used without imposing any higher-order structure, the document is formally modeled as an undifferentiated "bag of words."[4] For analysis, word instances or tokens are often stemmed for related roots (e.g., *pray*, *prayer*, and *prayed* all collapse to *pray*). Richer tokenizations also become possible by including frequent n-grams, or common word sequences of length *n* (e.g., bigrams *prayer meeting* and *gangsta rap* or trigrams *cup of tea* and *Central African Republic*), or skip-grams, n-grams with gaps of length *k* (e.g., 1-skip-bigrams *cup tea*

[4] These "bags" are not sets, but contain every instance of words used, and so retain frequency information.

Figure 1.1: Linguistic features for text analysis.
The first sentence from the Wikipedia entry on “Shooting of Trayvon Martin,” retrieved October 5, 2015, and the automated output (with punctuation tagging removed) extracted by Stanford CoreNLP (Manning et al. 2014). See http://stanfordnlp.github.io/CoreNLP/, or, to generate a user submitted example, http://nlp.stanford.edu:8080/corenlp/. Part-of-speech tags are from the Penn Treebank tag set (Santorini 1990): CD, cardinal date; DT, determiner; IN, preposition or subordinating conjunction; JJ, adjective; NN, noun (singular or mass); NNP, proper noun (singular); RB, adverb; VBD, verb, past tense. Directed linguistic dependencies are from the Stanford Dependencies representation (de Marneffe et al. 2014): advmod, adverbial modifier; amod, adjectival modifier; appos, appositional modifier; case, case-marking, preposition, or possessive; compound, noun compound modifier; det, determiner; dobj, direct object; nmod, noun modifier; nsubj, nominal subject; nummod, numeric modifier.


“On the night of February 26, 2012, in Sanford, Florida, George Zimmerman fatally shot Trayvon Martin, a 17-year-old African American high school student.”
Part-of-speech tags
IN DT NN IN NNP CD CD IN NNP NNP NNP NNP RB VBD NNP NNP DT
On the night of February 26 , 2012 , in Sanford , Florida , George Zimmerman fatally shot Trayvon Martin , a
JJ JJ JJ JJ NN NN
17-year-old African American high school student .
Named entities tags
Time Date Location Location Person Person
On the night of February 26 , 2012 , in Sanford , Florida , George Zimmerman fatally shot Trayvon Martin , a
Duration Miscellaneous
17-year-old African American high school student .
Coreferences tags
Mention Coref
On the night of February 26 , 2012 , in Sanford , Florida , George Zimmerman fatally shot Trayvon Martin ,
Coref Mention
a 17-year-old African American high school student .
Dependencies tags
nmod
nmod
case
det
nmod
case
nummod
nummod
case
compound
compound
nsubj
advmod
dobj
IN DT NN IN NNP CD CD IN NNP NNP NNP NNP RB VBD
On the night of February 26 , 2012 , in Sanford , Florida , George Zimmerman fatally shot
appos
det
amod
amod
amod
amod
compound
dobj
compound
NNP NNP DT JJ JJ JJ JJ NN NN
Trayvon Martin , a 17-year-old African American high school student .

and *Central Republic*). Each document can then be represented as a sparse vector of counts for each token in the vocabulary, with such counts often normalized by the number of tokens within document, or weighted to highlight the degree to which they distinguish the document.[5] While many text analyses, like sentiment classification, use only functions of the distributions of document words, the best performing systems include richer understandings of language structure (Hirschberg & Manning 2015).

Words refer to semantic *entities*, which may be referred to by many words (e.g., synonyms, pronouns). Co-reference is the process by which words are linked to this underlying entity, like the resolution of anaphora introduced by an author to vary his or her writing. The resolution of co-reference can improve document vectors. Moreover, in IE, some semantic entities are considered "named entities", pre-defined categories including names of persons, organizations, and locations. Named entity recognition and extraction are tasks involving the identification of these entities and extraction of details associated with their specific instances into a database (see Figure 1.1).

Syntax is the structure of words within sentences. There are many formal grammar-based approaches for uncovering sentence structure. The most common two involve parsing according to (1) phrase structure (or constituency) grammars, and (2) dependency grammars. The first seeks to decompose sentences into contiguous phrases. The second identifies a network of dependencies between words across the sentence.[6] Parsing sentences according to both approaches has become increasingly accurate in recent years. For example, in English text,

[5] The most common weighting scheme uses term frequency within a document, divided by the number of documents in the collection with at least one mention (tf.idf).

[6] Phrase or dependency representations may be more or less appropriate based on the degree to which word order is critical in the language or sublanguage of interest.

dependency parsing has come to exceed 95% accuracy by Google researchers (Coppola & Petrov 2015). Nevertheless, because of the computational complexity—and often inaccuracy—of parsing, local structure is often captured blindly through the use of n-grams and skip-grams. Alternatively, models can operate over the sequence of part-of-speech tags attached to the lexicon in order to robustly "chunk" a sentence into noun, verb and prepositional phrases. Social and computational analysts will often use word co-presence within a grammatical phrase, closeness in a dependency network, or proximity within a sequence of words to infer large-scale, semantically meaningful, associations between words. Social analysts may sometimes use linguistic structure more directly to extract unique data within textual claims. Consider Franzosi's network analysis of Subject-Verb-Object triples, like *cops* (Subject) *beat* (Verb) *protestors* (Object) (Franzosi 2004), which can increasingly be semi- or fully-automatically extracted as named entities related through parsed dependency relationships.

Not pictured in Figure 1.1, higher order document structure, or discourse, has also been productively used to analyze text. Work that examines word co-location within paragraphs (Lee & Martin 2015), author-created section headings (e.g., "Materials and Methods") or induced partitions of documents traced by lexical shifts (Hearst 1997) can provide valuable information about higher-order associations.

Another linguistic character that does not map unambiguously onto text is phonology, or the system of speech sounds. Distinguishing dialects through phonology can reveal distinct social worlds underlying spoken interaction (Rickford et al. 2015). For example, phonological cues from interviews performed with National Longitudinal Study of Youth respondents reveal how "ebonics", or urban, culturally-marked slang, accounts for wage gaps between respondents better than race (Grogger 2011).

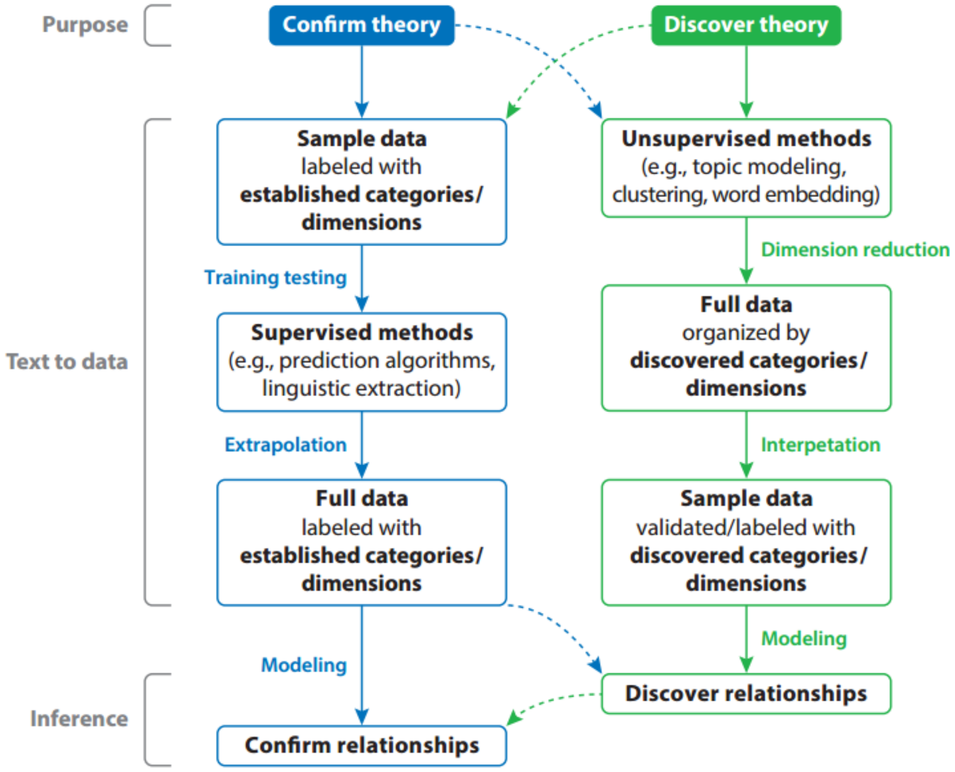
Improvements in all areas of NLP draw on a changing substrate of tools from probabilistic modeling, information theory, matrix factorization, and “deep learning” or multi-layer neural networks. Despite dramatic improvements in recent years, limitations remain. The most major is also a problem with sociological research: most NLP resources—like most sociological studies—are only available for high-resource languages such as English, Spanish and Chinese, and not low-resource languages like Bengali, Indonesian and Swahili, spoken by hundreds of millions of people (Hirschberg & Manning 2015). Another limitation relates to the lack of sophisticated models for higher-level linguistic discourse, such as how sentences relate to one another (Stymne et al. 2013) and aggregate into paragraphs and more or less effective arguments, although this is one of the targets of sociological text analyses (e.g., message complexity and popularity in Bail 2016).

While sociological analyses of text sometimes benefit directly from rich features of language induced with NLP, more often they take these features as inputs to ML models, which are themselves used to model issues of fundamental sociological significance, including collective attention, social relationships and socially relevant states.

Theoretical Purpose, Research Design, and Machine Learning Approaches

A sociological research project employing text analysis begins with either an impulse to evaluate pre-existing theory, or alternatively to explore, induce and discover theory related to the domain from which text was sampled. Figure 1.2 sketches how these distinct purposes can influence research design and shape the choice of ML methods used to create socially relevant data. Social theory further breaks down into (1) *concepts* that trace social entities, natural or imagined phenomena, and (2) *relationships* that link and structure them.[7] If concepts have already been identified in text, then the researcher will often use a supervised ML approach to extend these identifications to data beyond the analyst's capacity to reliably code. If concepts are yet to be discovered, then an unsupervised method will likely be draw upon to assist.

Figure 1.2: Text analysis and social theory.
The arrows trace the text analysis research pipeline, highlighting how different motivations—for confirming versus discovering theory—influence the choice of ML methods used to construct data from text relevant for sociological inference. Solid lines represent straightforward research pipelines, and dashed lines suggest research pathways that mix research motivations for confirmation and discovery. For example, a researcher might explore novel arrangements of established categories (e.g., sentences tagged with positive sentiment). Moreover, once new patterns are discovered in one corpus of text, they may be tested in another.


Purpose
Confirm theory
Discover theory
Text to data
Sample data labeled with established categories/ dimensions
Unsupervised methods (e.g., topic modeling, clustering, word embedding)
Dimension reduction
Training testing
Supervised methods (e.g., prediction algorithms, linguistic extraction)
Full data organized by discovered categories/ dimensions
Extrapolation
Interpetation
Full data labeled with established categories/ dimensions
Sample data validated/labeled with discovered categories/ dimensions
Modeling
Modeling
Inference
Discover relationships
Confirm relationships

[7] If the relationship is rendered a logical predicate, and the concepts its arguments, then together they formally comprise a theoretical claim. Consider the Marxist theoretical cartoon: if Capitalism ≡ *c*, and destroys ≡ *D*, then *cDc*.

Supervised and unsupervised approaches condition distinct research pipelines through which text is processed into socially relevant data.[8] With supervised methods, an analyst begins with a sample of text instances where concepts have been identified and coded by themselves or others. The concepts may be inherited from prior theorists, deduced from prior arguments, or discovered by an interpretive analyst in the process of coding. This sample is then divided into training and testing subsamples[9] and a supervised ML method draws on features associated with instances in the training sample to estimate a statistical model or tune an algorithm. The trained model or algorithm is then used to "predict" identified but unlabeled instances in the testing sample to evaluate its success. Success is typically measured with an IR metric that captures some balance of false positives and negatives (mistaken classifications and missed classifications) such as Precision, Recall or Area Under the receiver operator characteristic—ROC—Curve (AUC). If accuracy is not sufficient, more trained instances are identified by human coders, and the training and testing process is repeated to satisfaction. Finally the successful model or algorithm is used to extrapolate codes to unlabeled data. With data on established codes in hand, the analyst moves on to analyze hypothesized relationships between measured constructs with appropriate statistical models.

When unsupervised methods are used to discover novel categories or dimensions from text, the allocation of human effort is typically reversed. An automated ML model or algorithm is unleashed on the complete corpus of interest, furnishing new, discovered variables for

[8] This process of transforming unstructured data, like text, into structured data that is in turn leveraged to create new forms of value has sometimes been termed "datafication"(Schutt & O'Neil 2013).

[9] This division can often be dynamic and changing, as in a tenfold cross-validation design, where the data is split into ten parts, and the model is successively trained on nine parts and tested on the tenth, cycling through each split.

subsequent analysis.[10] Such models are given clues about the patterns or rules they should be learning (Jurafsky & Martin 2000, p. 117) and rest on assumptions about the underlying structural properties of the data, whether algebraic, combinatorial, or probabilistic (Jordan & Mitchell 2015). The algorithms learn distinguishing characteristics such as distributional patterns or clustering properties (Clark et al. 2010). Occasionally the resulting unsupervised data structures are automatically labeled and trusted, but more often analysts formally or informally sample, peruse and interpret them. For example, topics produced by a probabilistic topic model estimated on a corpus are scrutinized, then explicated and hand-labeled for easy description and reference. As with supervised models, the data that results is often subsequently used to discover or confirm theoretically significant relationships with an appropriate statistical model.

In sociological analyses, text-based variables derived from supervised or unsupervised methods usually take an explanatory role as independent variables that predict an established dependent variable from outside text. For example, Goldberg et al. (2015) extract the degree of an employee's cultural embeddedness within a firm from text, and then use it to predict individual performance ratings and tenure. In the context of this usage, we now define some of the most recent and promising supervised and unsupervised approaches for sociological text analysis.

Supervised methods begin with a training sample of text, tagged with expert-defined codes to identify categories of interest, like positive sentiment, liberal ideology, or mention of a particular social movement strategy. This selectively tagged text furnishes positive and negative

[10] Unsupervised models need not "discover" new variables. They may, instead be used to discover known categories, in which case they are built on training data and evaluated on testing data as with supervised models. Such is the case with unsupervised syntactic parsing models, where the outcomes are known. (Nelson 2015) is a sociological case in which unsupervised topic models were used to identify established differences between the discourse of feminist organizations in Chicago and New York over time.

examples for a supervised model to distinguish. Supervised models include linear and logistic regression, but with thousands or tens of thousands of independent variables corresponding to text features like the word frequency vectors described above (Joshi et al. 2010). Given the high dimensionality of text data, it is not always possible to efficiently estimate these models without simplifying text variables and reducing their dimensionality. Most directly, social science applications have analyzed contingency tables to identify distinguishing n-grams from positive and negative examples (Gentzkow & Shapiro 2010; Laver et al. 2003), which are subsequently included as predictors in regression models to identify sentiment, policy positions or ideological slant.[11] Principal components regression (Massy 1965) and supervised LDA topic models (Mcauliffe & Blei 2008), have also been used to reduce the text dimensionality by deriving components or topics, subsequently used as predictors in a regression. An integrated approach with stronger performance is multinomial inverse regression, a two-stage estimation approach in which linguistic features are first regressed on some function of the category of interest (e.g., positive sentiment), then selectively included in a forward-regression without losing responsiveness to the predicted category (Taddy 2013).[12]

Alternative approaches use a full or partial complement of word frequencies and related linguistic features in a range of other ML classification algorithms, including k-nearest neighbor analysis, naïve Bayes estimation, support vector machines (SVMs), maximum entropy classifiers, deep learning, decision trees and ensemble techniques that combine the judgment of multiple approaches. Some of these approaches maximize interpretability (e.g., pruned decision

[11] This approach hearkens to classic approaches that relied simply on weighted counts for a pre-defined list of terms (Loughran & McDonald 2011; Tetlock 2007).

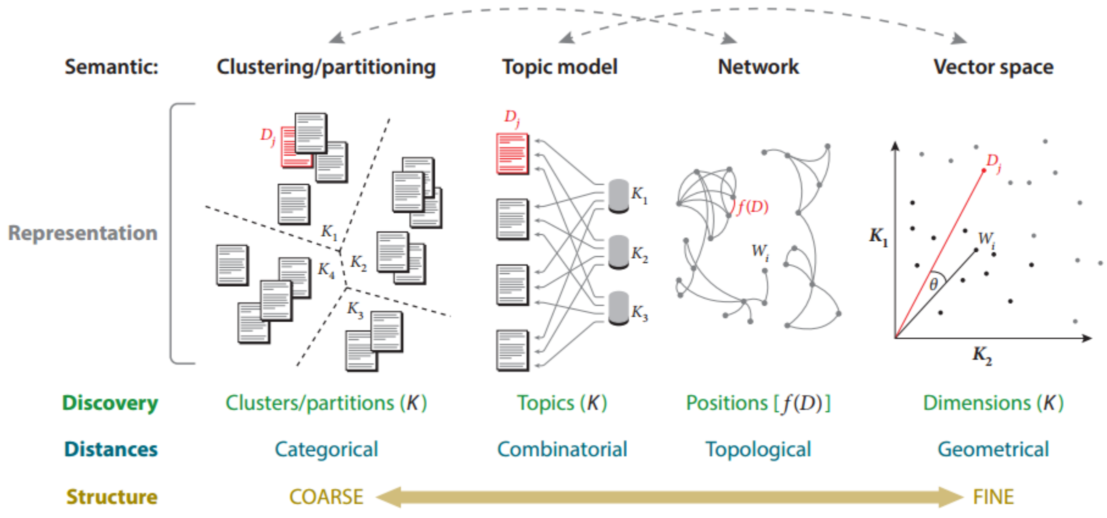
[12] This is similar to sparse regression approaches like the least absolute shrinkage and selection operator or LASSO, which minimizes the usual sum of squared errors, but with a bound on the sum of coefficient absolute values to select a sparse subset of high-signal predictors (Tibshirani 1996).

trees), accuracy (e.g., deep learning and ensemble techniques) or speed (e.g., SVMs).[13] All of these approaches can be generalized to content beyond text, including audio, images, and video with deep learning approaches recently becoming dominant (Hinton et al. 2006; Hirschberg & Manning 2015).

Unsupervised methods begin with a corpus of unannotated text, then discover and represent novel structures for interpretation. We highlight four of the most common: clustering, network analysis, topic modeling, and vector space embedding, each illustrated in Figure 1.3 along with attributes of the data representations they produce. Clustering is typically used to discover

discover

Figure 1.3: Properties of four common semantic representations.
Four of the most widely used semantic representations of text, linked to the most frequently deployed unsupervised machine learning approaches, are clustering, topic modeling, semantic network induction, and vector space word embedding. In these representations, D represents documents, W represents words within those documents, and K represents discovered semantic structures. Document clusters, in which similar texts are typically grouped by shared words, closely relate to semantic networks, in which similar words often link as a function of the documents in which they co-occur, f(D). Topic models, which represent documents as sparse mixtures of induced topics, are closely related to vector space word embeddings, which define documents, words, and phrases as dense mixtures of induced dimensions, or vectors plotted in a space anchored by those dimensions.



[13] Examples can be found in (Pang & Lee 2008; Srivastava & Sahami 2009; Yu et al. 2008).

coarse-grained, categorical groupings of documents through their words, while network analysis has typically been used to identify fine-grained topological positions of words and their underlying entities across documents. Topic modeling has been used to coarsely describe documents as sparse combinations of latent topics, while word embedding models spread words and documents across high-dimensional spaces from which semantic distances can be calculated.

In clustering documents, similar texts are hierarchically grouped, dissimilar texts hierarchically divided, or some function of intra-cluster similarity and inter-cluster difference is maximized. Such algorithms are often performed on document vectors including weighted words. Document cluster assignment can be exclusive and "hard", as in all hierarchical models, or soft, allowing for degrees of membership. Different clustering rules produce different clusters, which in turn reveal different social games from the data (Grimmer & King 2011). For example, King and Grimmer cluster U.S. Congressional press releases to reveal a common, but previously unnoticed genre of "partisan taunting" in which one political office-holder berates another to highlight their own positions or justify political action.

A semantic network approach links words or phrases co-located within documents, sentences, clauses, and dependency parse trees. This approach can be considered the fine-grained corollary of document clustering. This unsupervised and largely model-free approach can reveal the fine structure of cognitive and cultural association between entities through calculation of their network analytical positions, such as word centrality, influence, structural equivalence and constraint (Atteveldt et al. 2008; Carley 1993; Carley & Kaufer 1993; Schank et al. 1973; van Atteveldt 2008).[14] For example, Carley used semantic networks to trace shifts in culture, like the evolution of robots from alien monsters to sympathetic companions in fiction (1994), and

[14] These networks can also be partitioned, a graph-based approach to clustering.

Corman has used aggregated betweenness centrality across noun phrases to highlight terms that channel meaning (2002).

Topic modeling is a recent and increasingly common approach to discover semantically cohesive “topics” and their combination across document collections. Topic models are an influential class of generative, Bayesian probabilistic models that model documents as draws from a set of induced topics. Formally, each topic is a latent multinomial variable, tracing a distribution over all words in the corpus vocabulary. The first topic model was titled latent Dirichlet allocation (LDA), because a Dirichlet distribution was used to draw per-document topic distributions in the first stage of the model (Blei 2012; Blei et al. 2003). This distribution is typically tuned to minimize the mixture of topics describing any particular document in the collection. Estimated topics become objects of interpretation for the sociologist, who uses them to describe document collections and trace collective attention and reasoning for the culture, organization or community that authored them. A recent special issue of *Poetics* was entirely devoted to topic modeling in social and cultural analysis (Mohr & Bogdanov 2013; McFarland et al. 2013b). Many topic model variants have been proposed, including those that account for local word-order and syntactic dependencies (Griffiths et al. 2005; Wallach 2006), explicit document dependencies like citations or hyperlinks (Chang & Blei 2010), the temporal order of documents and evolving topics (Blei & Lafferty 2006).

Word embedding models construct an efficient set of dimensions from a document collection, in which all documents and words can be projected. This approach is conceptually related to topic modeling, except that because stable semantic distances and not descriptions are the goal, documents are not sparsely embedded over dimensions, and dimensions are not sparsely embedded over words, as they typically are in topic models. Older approaches to Latent

Semantic Analysis (LSA) used singular value decomposition on the document-word matrix to identify informative dimensions (the singular values) in which words and documents could be projected. Distances are often calculated between the angles of these vectors to assess semantic distance irrespective of document size. For example, the cosine distance between a single word (e.g., *America*), a one line search engine query (e.g., *United States of America policies laws*), and an entire website (*usa.gov*) could be very small suggesting close semantic relation. More recent approaches use neural networks on a large set of linguistic features, including not only words and n-grams, but skip-grams, to encode a wide range of syntactic and semantic information. The most prominent of these is word2vec developed by a team of Google engineers. Word2vec produces word vectors that perform well on human analogical reasoning tests requiring substantial human semantic understanding. Consider the question, "man is to king as woman is to ___?" (queen). In the high dimensional geometry of these vector space models, with 100-500 induced dimensions, the vector for *man* plus the vector for *king* minus the vector for *woman* is closest to the vector for *queen* (Mikolov et al. 2013b; Pennington et al. 2014).[15] A recent entrant in this model class is Global Vectors for Word Representation (GloVe), which combines global matrix factorization, to capture word associations across each document, with the rich, local context windows described above (Pennington et al. 2014). Once learned, vectors from these approaches can reveal the fine structure of words within a cultural world traced by its texts. Semantically similar words and documents will show up in the same angular region of the high dimensional document space. Like the topics from topic models, dimensions and distances from

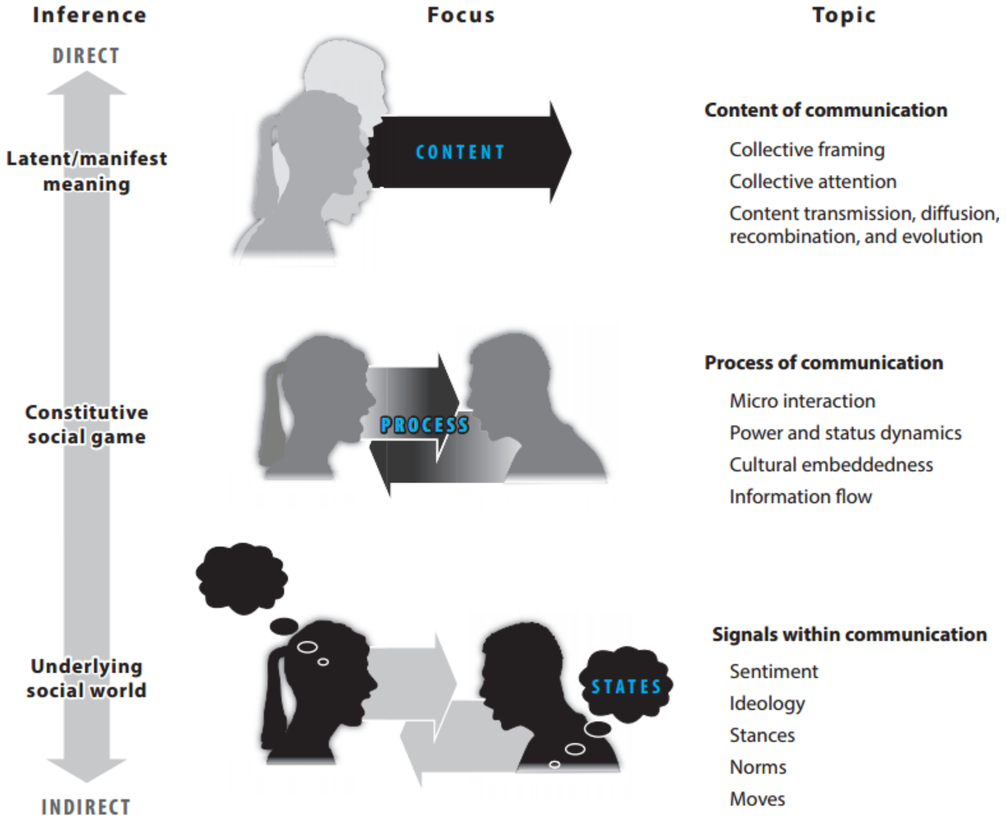
[15] Such models perform between 70% and 80% accuracy on analogy tests. When we recently used the model, trained on a large Google News corpus, to probe cultural associations, we found many other associations suggestive of its ability to probe cultural and cross-cultural fields, like *black* + *hiphop* – *Mexican* = *norteño*.

trained vector spaces can also be deployed in a wide variety of supervised tasks including sentiment analysis, document classification, and entity recognition.

Mining Text for Social Theory

Many recent empirical articles and conference proceedings apply these NLP and ML methods to address questions central to sociological concerns. These articles sort themselves by the layer of communication on which they focus, and the depth of inferences they make about the social world (see Figure 1.4). First, we review the substantial quantity of research devoted to generating knowledge from the manifest and latent *content* of communication in text. This research uses meaning-filled words, topics and topic-shifts to address patterns of collective framing and attention, but also explores how society "thinks" by tracing the transmission, diffusion, recombination and evolution of content. Second, an emerging stream of research has begun to study social relationships by analyzing the *process* of communication. These papers use dynamic patterns of linguistic mimicry and synchrony to trace the deep and often hidden dynamics of social interaction underlying communication and the information that passes across them. Papers here focus on micro-interaction, power and status dynamics, cultural embeddedness in communication, and information flow. Finally, there is burgeoning interest in using heterogeneous linguistic *signals* within communication to analyze social identities, states, roles and moves. These papers attempt to access deep information within text about hidden elements of the social game being played and the social world beneath it. This work focuses on tracing sentiment, ideology, stances, and norms, but also identifying actor roles and predicting future moves. We also note that some productive computational text analysis seeks to make no inferences, but simply draw on user-generated codes to bootstrap expanded samples of similar passages meriting qualitative investigation and inference. Such approaches directly extend qualitative text analysis by using NLP and ML to index enormous, unreadable libraries.[16]

Figure 1.4: Social Inference from Communication.
Computational text analysis articles arrayed according to the depth of inferences they make about the social world, with foci ranging from (1) collective attention, framing, and "thinking" through the manifest and latent *content* of communication; (2) social relationships through analysis of the *process* of communication; and (3) social identities, states, roles and moves through linguistic *signals* embedded in communication.



Collective attention and reasoning through the content of communication. These papers most directly extend the classical concerns of content analysis (Berelson & Lazarsfeld

[16] Tangherlini and Leonard (2013) used Google books to sample passages with familiar themes (e.g., references to Darwin and evolution) in the vast archive of unread works. Shahaf et al. (2012) take a higher-level approach by using metrics of influence, coverage, and connectivity from the scientific literature to create structured summaries, or "metro maps", of information for expert perusal. These maps helped users find better and more seminal papers, and perform fewer queries than those using standard web search.

1948; Woodward 1934), initially using newspaper content, by focusing directly on cultural forms within communication—frames, issues and topics. This work identifies the structure and dynamics of attention, agreement and search across settings ranging from small groups and organizations to vast, far flung publics like the Twitter-sphere. Patterns of content can be stable or changing; unified, fragmented or polarized. Studies tracing the dynamics of content trace how social collectives process information—how they collectively "think"—by tracing the growth, diffusion, mutation, recombination, and extinction of issues and topics.

Content has been used to directly investigate the structure of meaning in fields ranging from institutional logics, politics and economics to culture and idling conversation. In a multi-method study, Nelson traces the institutional logics of women's organizations in Chicago and New York City during the early and middle 20th Century (Nelson 2015). By applying LDA topic modeling and qualitative in-depth readings to a corpus of women's movement organization publications, she shows that city-specific political logics persist over time. The project demonstrates that "organizations institutionalize and embody local cultures that are then drawn on over time as new organizations are formed, and partially determine the distribution of resources to new organizations, producing within-city cultural continuities" (Nelson 2015). Topic models allowed her to inductively discover continuities in these logics and incorporate the findings into a historically informed understanding of the U.S. women's movement, while extending theory on how the local institutionalization of cognitive structures persist over time.

Consider recent work on political frames held by legislators, media organizations, consumers and constituents (Grimmer & King 2011; Grimmer & Stewart 2013). Investigating senatorial press releases from 2005 to 2007, Grimmer (2013) uses a topic model extension (Grimmer 2010) to simultaneously estimate the topic of each press release, the proportion of releases senators produced on each topic, and the category into which senators fell each year in office. This revealed how legislators present themselves in a way that reflects their political alignment with constituencies. Politically aligned senators stake out political positions, while misaligned senators avoid controversial positions and instead claim credit for appropriations to their districts. Gentzkow and Shapiro explore the link between politics and the market (2010) by developing a measure of media slant using phrases from the Congressional Record that statistically distinguish Democrat and Republican speakers. This allows them to estimate the degree to which hundreds of American news media outlets echo Democrat vs. Republican Congressional voices. When they combine slant with circulation data, Gentzkow and Shapiro find that the ideology of potential newspaper readers and not owners match the newspaper's slant, suggesting an incentive to "tailor…slant to the ideological predispositions of consumers" (2010). In related work on the effect of Facebook's News Feed on users' consumption of discordant ideological content, Bakshy et al. directly measured Facebook users' expressions of political commitment, examined the relative influence of user choice and Facebook's algorithm at limiting exposure (2015). They found that while News Feed slightly limits exposure, individual choices to avoid discordant content play a much larger role.

Other recent work structures content according to the characteristics of those that produce it. For example, Jockers and Mimno (2013) studied works of 19th century British, American, and Irish literature, demonstrating how factors such as author's gender, nationality, and precise historical period of publication affected fluctuations in themes and word choices used to articulate them.

Still other research uses computation to trace the allocation of attention across societal domains. Bail (2012) uses plagiarism detection software to compare how national newspapers and television news stations distributed their attention across press releases about Islam by civil society organizations in the wake of the September 11th attacks. He finds that the mass media paid more attention to fringe organizations. This then realigned organizational networks and shifted discourse surrounding terrorism. Similarly, Bonilla and Grimmer (2013) use LDA topic models to document how newspapers and nightly newscasts from major network stations allocated their attention to terrorism after elevation of the U.S. government's color coded alert system.

The coherence and diversity of content has also been explored. In the context of political campaigns, Livne et al. (2011) analyze tweets from federal candidates in the 2010 midterm elections to estimate content cohesiveness. They found conservative candidates portrayed the most coherent message, with Tea party candidates displaying surprising topical and linguistic cohesiveness despite their lack of formal organization. Tsur et al. (2015) use press releases from U.S. representatives to investigate political framing and agenda setting campaigns. Using LDA topic modeling and autoregressive-distributed-lag models, they find significant differences between the framing strategies of Democrats and Republicans. In the context of social movements, Bail (2014) investigates how advocacy organizations for organ donation produce different discourses in their appeal to multiple audiences. By developing a theory of cultural carrying capacity and putting it to work through structural topic modeling, Bail finds an inverted-U shaped relationship between a campaign's message diversity and social media endorsements, comments, and shares. Diverse content (e.g. social media messages that discuss religion, sports, or science) leads to more public engagement, but only up to a point, after which campaigns appear incoherent, lacking shared purpose or collective identity.

A large corpus of recent work examines shifts in content over history to identify changes in the social world underlying it. Consider Michel et al. (2011) who descriptively follow shifts in the usage and meaning of a wide range of terms and phrases from millions of books. They term their approach culturomics or "the application of high throughput data collection and analysis to the study of human culture", and apply it to detect the rise of censorship in WWII Germany, the trajectory of fame, and shifting conventions of gender. In the political domain, Rule et al. (2015) use community detection algorithms on State of the Union transcripts from 1790 to 2014 to trace the emergence of modern political discourse through collocated terms. Likewise, Klingenstein et al. (2014) empirically document the emergence of Elias' "civilizing process" through analysis of transcripts from the criminal proceedings in London's Central Criminal Court, the Old Bailey. Using the information theoretic Jensen-Shannon divergence, they found that from 1760 to 1910, discourse around violent and non-violent crimes underwent a gradual, but massive differentiation. Finally, Miller (2013) uses LDA topic modeling on the Qing Veritable Records (a collection of Chinese documents important to an emperor's reign) to model typologies of violence held by government administrators. This typology provides insight into how different epochs understood violence and elucidates changing "crime rates" during the 18th and 19th centuries.

DiMaggio et al. (2013) trace discursive change by inducing topics from newspaper articles between 1986 and 1997 that discussed the National Endowment for the Arts (NEA). They found the tone of news coverage about the NEA shifts after the election of George H.W. Bush, from celebration to controversy and negativity. Marshall (2013) explores the topical nature and shifts within the discipline of demography in England and France. By tracing the prevalence of topics in 3,458 articles from 1946 to 2005, she delineates how research trends from each country reflect the cultural and institutional differences that shaped each country's understanding of fertility decline. Mohr et al. (2013) take a more structured approach to temporal shifts. Using named entities to identify actors, part-of-speech tagging to locate actions, and topics to capture scenes of action, the authors operationalize Kenneth Burke's "grammar of motives" across 11 national security documents produced by the U.S. government between 1990 and 2010. Mapping rhetorical forms to structural properties, this work offers a suggestive approximation of deeper, more interpretive analysis regarding how the state legitimates its policies.

Attention and information diffusion follow stable, recurring patterns within many social games. Instant messaging conversations, fields of news production, and scientific disciplines each processes information in characteristic ways. This *collective reasoning* results in stable patterns of attention, diffusion, mutation and combination of content. Within mass and social media, Leskovec et al. (2009) have analyzed the temporal dynamics of the recurring news cycle on a large dataset of news and social media sites. Tan et al. (2014) document the effect of message wording on the diffusion of tweets by analyzing millions of paired tweets. Finally, Cheng et al. (2014) trace the nature of photo re-share cascades on Facebook (over 150 thousand photos), using caption and photo features as predictors. In science, Kuhn et al. (2014) have traced the diffusion of scientific memes, n-grams that propagate along the scientific citation graph, by analyzing published science in physics, biomedicine, and a broad sample of science and scholarship. Foulds and Smyth (2013) have similarly analyzed the proceedings of conferences to create a measure of topical influence that tracks the degree to which cited articles spread their topics to citing articles.

Other studies analyze the evolution of content, or how it transforms as it travels through communication channels. Much of this work looks at specific shifts in word meanings or the degraded fidelity of messages as they spread. Kulkarni et al. (2015) use word embedding models built from Twitter, Amazon movie reviews, and a hundred years of Google book ngrams to find change points in word meaning and usage. Following the word “gay” across the 20th Century, they trace its trajectory from the neighborhood of “dapper” to that of “lesbian”. Adamic and colleagues (Adamic et al. 2014; Simmons et al. 2011b) analyze the dissemination of memes replicated millions of times on Facebook, discovering regularities governing the evolution of socially shared information. Simmons et al. (2011b) use the Meme Tracker dataset to find that mutations in quotations depend primarily on the authority of the copied source and nature of the quoting website. A paper by Gross (2014) explores the evolution of design logo content across an individual’s design history to measure the creative effort behind each logo, and identify incentives for creativity within design competitions. This paper uses image instead of textual content, but its edit-distance measure of difference between images is conceptually identical to Adamic et al.’s measure of difference between memes (2014). This research suggests how shifts in content can be used to uncover deep cultural change relevant to many sociological concerns, from changes in gender roles to shifts in the economy and innovation.

Another class of articles explores the characteristic process by which textual elements are combined over time to suggest how collectives think, search and discover, and how they could be redesigned to do it more efficiently. In the context of financial policymaking, Fligstein et al. (2014) investigated the lack of awareness at the Federal Open Market Committee (FOMC) of the impending economic meltdown in 2008. Through topic modeling, they show that the Federal Reserve’s primary analytic framework of macroeconomic theory prevented the FOMC from connecting the disparate forces of the crisis—the housing market, subprime mortgage market, and financial instruments used to package mortgages into securities—into a comprehensive picture. The common assumptions of those in charge at the Federal Reserve led to an inability to combine categories in the requisite ways to make sense of the crisis.

Other work tracing how collectives think has been carried out in science, which leaves a particularly detailed published trace of its content. Shi et al. use random walks along the network of scientific content from millions of biomedical abstracts from the MEDLINE corpus to reveal typical patterns of discovery (2014). Foster et al. (2015) build on this by creating a network typology of discovery that reveals strong institutional pressures for scientists to exploit prior knowledge through incremental advance rather than explore the explosively expanding set of new opportunities. They find that rewards for high-risk innovation are not sufficient to compensate for the risk of not publishing, leading to few high-risk innovations. Rzhetsky et al. (2015) build a generative probabilistic network model to trace how molecules are typically combined in biomedicine and chemistry research, then estimate it with extracted content on molecules mentioned in millions of papers and patents over the latter half of the 20th Century. They then discovered optimal strategies through simulation and compared them with historical modes of collective discovery to identify inefficiencies associated with science as currently organized. These inefficiencies trace institutional incentives, like tenure, which value sustained incremental productivity by an individual researcher over risky, collective advance.

These insights can also generate prescriptions, like Spangler et al. (2014) who use a similar approach, combining entity detection and graph-based information diffusion models to identify potential but untested relationships among scientific entities. They then examine these hypotheses through laboratory experiments and discover novel relationships implied by the corpus. Research that explores the process through which content is recombined to generate innovation has tended to focus on domains like molecular biology where nouns and verbs are well-behaved (e.g., chemicals and reactions). Nevertheless, combined with robust approaches to dimension reduction, this work suggests a mode of analysis for precisely identifying and evaluating how institutions think relevant to the study of organizations, communities, cultures and social movements. Recent movements like Occupy Wall Street, Black Lives Matter, and Fair Trade are all composed of evolving claims and shifting attention that could be better understood through the extraction of content with NLP tools at scale.

Social relationships through the process of communication. Beyond cultural forms, burgeoning attention in computational text analysis is now being paid to how such tools can be used to explore social relationships as they unfold through the process of communication. These issues have been core concerns of qualitative approaches including ethnomethodology (Garfinkel 1967), studies of the interaction order (Goffman 2005), and conversation analysis (CA) (Schegloff 1992). Research using NLP and ML to investigate social relationships and the social games underlying communication is beginning to uncover deep regularities in interaction and opening up exciting new areas of research. This research deals with communicative synchrony, the flow of information that results, and what this reveals about the relative positions of interaction partners and their outcomes for individuals, dyads and groups. Data typically involve turn-taking patterns, the interactive pattern of words, phrases and higher-order topics, and the movement of information through the ongoing flow of communication.

Recent computational research documents how social actors match one another's linguistic style, but unequally.[17] The imbalance reveals deep insight into positions within the

[17] In Adam Smith's *Theory of Moral Sentiments*, he promoted a "propriety" of both social and linguistic action that foregrounded sympathy, or taking the other's perspective for achieving mutual correspondence and social harmony (Dascal 2006; Smith 1759, sections 1.1.1.3, 1.1.1.5). Smith mandated that sympathetic effort be equitably divided between interacting parties, for it would be both improper and inefficient for either to lay upon the other "the whole burden" (Dascal 2006; Smith 1759, section 7.4.28).

social game constitutive of communication.[18] For example, Danescu-Niculescu-Mizil et al. (2012) use the micro-dynamics of language coordination to trace shifting power differences between interaction partners by showing that lower status parties to a conversation engage in conscious or unconscious mimicry of the distribution of function-words (e.g. articles, prepositions, personal pronouns) expressed by those with higher status. In their analysis of conversational exchanges on Wikipedia forums, they find that low-power contributors end up coordinating much more consistently to the linguistic style of high-power administrators, and that this shifts as editors are voted into and out of administrative positions. They also apply this manner of tracing linguistic style coordination to oral arguments in the U.S. Supreme Court. Low-power lawyers coordinate much more consistently with high-power justices than the reverse.

Even though those in favorable positions may be less likely to match the style of others, inability or unwillingness to match another's style does not improve one's position. Recent computational work using a similar design to that above demonstrates through U.S. presidential debate and negotiation transcripts that polling increases for presidential candidates and third-party evaluations favor presidents who mirror the language of their interlocutors more (Romero et al. 2015). This is likely because matching an opponent's style translates one's argument for ease of understanding and reflects the ability—and flexibility—to take the other's perspective. This matches with findings from computational analysis of recorded negotiations that more communicative mirroring is positively associated with favorable negotiation outcomes (Pentland 2014).

[18] Later versions of Adam Smith's *Theory of Moral Sentiments* included an appendix, "Considerations concerning the first formation of languages." Scholarship shows how his ethics of social sympathy in the body of *TMS* has a direct corollary to how he describes linguistic interaction (Dascal 2006).

Language style matching is not the only way to obtain domain-independent traces of power. Prabhakaran et al. (2014b) use indicators of overt displays of power (ODP) that place constraints on the recipient ("come to my office now") to identify power dynamics within the corporate email communications of top Enron executives (Prabhakaran et al. 2012). Indicators were obtained using a supervised model trained on manual annotations, and revealed that male superiors used significantly more ODP's compared to male and female subordinates, and that female superiors used the least ODP's of all. Another approach to power measurement uses message control. Prabhakaran and colleagues analyzed U.S. presidential primary debates to identify speaker turns where conversation changed from one topic to another (2014b). Higher-powered candidates—those posting higher poll numbers—were less likely to shift topics during the course of the debate, suggesting that control of the conversation is a consequence of electoral popularity.

McFarland et al. (2013) explore moves that take place during the game of courtship by analyzing audio and textual content from thousands of speed dating encounters. Marshaling both acoustic and linguistic data from the transcribed conversations, they derived measures related to emotional intensification (prosodic attributes such as pitch, loudness, and rate of speech) and conversational synchronization (lexical, syntactic, pragmatic, and interactional attributes such as interactional targeting, interpersonal alignment, and situational alignment). They found that mutual excitement is related to social bonding and the selection of a speed-dating partner for further contact, but this was contingent on gender. For men, excitement is expressed through laughter and variance in volume, while for women it is expressed through the raising and varying of vocal pitch. Further, participants felt a connection when men expressed sympathy and gratitude for their dates, and women engaged the situation and targeted themselves as a subject

("I"). It was unclear how much of this particular courtship game resulted from immutable differences in U.S. gender identities versus the choreography of the game itself, in which men were instructed to physically rotate and women stayed still, possibly giving them the upper hand in interaction.

Social relationships have also been explored in work on social and cultural embeddedness (Pachucki & Breiger 2010). Danescu-Niculescu-Mizil and colleagues (2013) explore this by using a decade of reviews from two review communities, to discover a cultural life cycle through which users come to learn online community norms of expression. New members enter the community, unwittingly introducing and perpetuating linguistic innovations, but eventually synchronizing with the community. Subsequently, their language stabilizes and becomes rigid, but community language norms continue to evolve and they eventually fall out of sync. This disconnect with community language predicts their disengagement and ultimate exit, prompting the paper's title "No Country for Old Members".[19] Goldberg et al. (2015) investigate cultural and structural embeddedness and their joint influence on individual attainment within an organization. Analyzing millions of emails from a medium-sized technology firm, they operationalize a measure of cultural embeddedness based on the degree to which the language people use is more or less similar to the within-organization emails they receive in a given month. This measure of cultural embeddedness is then analyzed along with their structural embeddedness—position in the email network—to find that structural and cultural embeddedness are inversely related with respect to community advance. Brokers do better when

[19] In a nontextual, but frequency-based analysis of communication embeddedness, Saavedra et al look at the pattern of instant messaging activity among traders in a day-trading firm and suggest that instant messaging patterns among their individual networks enable them to trade synchronously, which in turn decreases their likelihood of losing money at the end of the day (Saavedra et al. 2011).

they demonstrate higher cultural fit, while individuals in structurally cohesive positions do better when they perform cultural distinction.

At the field level, Vilhena et al. (2014) analyze the nature of communication within and between scientific disciplines by capturing the distribution of phrase frequencies in articles as well as the citation linkages between them from the JSTOR corpus (1.5 million scientific articles). The cultural space is mapped in terms of the communicative burden placed on interacting individuals, measured as the ratio of entropy and cross-entropy rates between members of one community and those of another. This means that individuals operating under very distinct phrase distributions—e.g., sociology and molecular biology—will expend much more effort understanding one another than those interacting against a backdrop of similar phrase distributions—e.g., economics and political science. Knowing the phrase frequencies of different subfields allows for the tracing of entire cultural spaces, which can then be mapped onto interactional structures (such as citation flows), leading to a topographical landscape rich in insights.

Patterns in the process of communication alter how information flows through the relationships facilitating it. Aral and Van Alstyne (2011) use hundreds of thousands of email messages at an executive recruiting firm to understand how novel information flows through the network. They measured bandwidth as the monthly volume of communication between individuals, and novelty as the introduction of distinct new content in a vector space model of email content. They found that as recruiters communicated across structural holes, they received more diverse information, but forewent communication bandwidth, reducing the overall volume of novel information received. In this way, social relationships are traced not only through webs of interaction, but also the process and content of the communication itself.

Alex ‘Sandy’ Pentland, his Human Dynamics Laboratory within MIT’s Media lab, and collaborators have used wearable devices like sociometric badges to reach beyond text and focus on nonlinguistic features of human communication, like patterns of turn-taking, tone of voice, facial movement and gesture to measure properties like activity, engagement, and mirroring within conversation. They have then used these qualities to predict outcomes. For example, rapid, even patterns of turn-taking within groups is associated with greater problem-solving success (Woolley et al. 2010a). Moreover, the presence of group conversations in which members face one another, side conversations and energetic engagement outside meetings are all associated with greater satisfaction and collective group performance (Pentland 2012a).

Social encounters “with differentially empowered individuals, complementary parts, and mixed motives” are common across many social games (McFarland et al. 2013a). While power and status dynamics may easily be observed through markers such as professions or institutional affiliations, understanding their subtle and various interaction markers through NLP and ML tools is adding detail in the form of social processes and mechanisms to classic sociological arguments, while discovering new ones. Collectively, articles on power dynamics (Danescu-Niculescu-Mizil et al. 2012; Prabhakaran et al. 2014a,b) present strong evidence for domain-independent methods for inferring status relationships within communication processes. We argue that these methods can be usefully applied to a wide range of sociological inquiry where power and status dynamics are of special interest. Contexts in which race, class, gender, occupation and other identities condition differences in the balance of interaction include criminal courts, on-line discussion forums, town-hall meetings, and anywhere else with readily available textual data stores. Many of these contexts may also generate large stores of non-digital information, which can now be transferred to digital form through Optical Character Recognition and speech-to-text technology, allowing scholars to trace the historical evolution of such relationships. ML approaches to the automatic classification of interactional variables from audio and video data could dramatically expand the interactional data available to Conversation Analysts (Stivers et al. 2009).

Social states through heterogeneous signals within communication. Linguistic communication exhibits regularities that serve as signals tracing deeper aspects of the social world underlying it. This section reviews work that makes inferences about human and collective states, which underlie and conceptually precede both content and interaction. The research outlined below uses statistical models and ML algorithms to predict phenomena beneath human communication at or beyond the performance of human experts. Text and associated communication data are used to identify internal human states like sentiment and preference, social roles and stances behind communicated utterances, and dispositions that predict future strategic moves.

Human and collective states reflect their condition at a given time and place. States include sentiment, preference, ideology, stance (for or against), norms, uncertainty, and disposition to perform a predicted action in the future. Sentiment analysis has been at the center of computational content analysis since Philip Stone's General Inquirer System used dictionary-based classifiers to capture lexical traces of positive, negative and neutral affect from speeches and news (Stone et al. 1966). Scholars have continued to make progress in reliably extracting sentiment from text, with existing methods that deploy neural network models over dependency trees achieving greater than 85% accuracy (Socher et al. 2013). These approaches can capture the effect of contrastive conjunctions (e.g., "but", "however") as well as negation (e.g., "not", "neither") and its scope at various levels within grammar trees for both positive and negative phrases.

Research has used sentiment analysis at both individual and collective levels. At the individual level, Sudhof et al. (2014) model dependency paths of human emotional states within the context of product reviews, finding that reviewers are swayed by sentiment from prior reviews. Similarly, Kramer et al. (2014) document large-scale emotional contagion in their controversial study using Facebook to demonstrate how small changes in the visibility of positive and negative content within a user's News Feed influenced that user to mirror this affect in their own posts. Taking these insights to the clinic, Resnik et al. (2013) used dictionaries, supervised classification, and topic modeling to aid in the clinical assessment of neuroticism and depression through sentiment identification. At the macro level, Golder and Macy find globally characteristic patterns of mood across day, week and season expressed on Twitter (2011), and several studies have looked at the effect that collective sentiment can have on stock market returns (Bollen et al. 2011; Nguyen & Shirai 2015). Other work has examined the recursive effect that events, captured by news volume, have on collective sentiment (Tsytsarau et al. 2014).

Tracing ideology has become an active area of research predicting human states. Iyyer et al. (2014) apply a deep learning framework to infer the political position of sentences and Sim et al. (2013) learn an ideological space from a corpus of explicitly ideological books and then use a probabilistic model to predict ideological valence from politician speeches. Jelveh et al. (2014) correctly predict the ideological leanings of professional economists through a supervised ensemble n-gram model applied to their research papers.

Stance (i.e., for or against) and disagreement have been analyzed in the context of online debates, with Sridhar et al. (2015) evaluating author and post-level stances within online debates, and Hasan and Ng (2014) going deeper to analyze not only stance, but reasons behind it. At the level of markets, Baker et al. (2013) have used automated text search for ten of the largest newspapers to create a measure of national-level economic uncertainty and trace it through historical news to identify its influence on capital markets. While most psychological and social states have been examined at the individual level, research has begun to trace shared or collective states that underlie communicated content, which opens up new opportunities for the sociology of emotion, culture and knowledge.

Clues within communication have also been used to predict strategic motivations and their influence on future moves in a variety of social games. One study used the online game *Diplomacy,* in which individuals engage in dyadic exchanges to form and dissolve alliances, to trace the linguistic harbingers of betrayal (Niculae et al. 2015). When the linguistic balance between partners changes by becoming more positive, polite, or focused on the future, betrayal follows soon thereafter. Similarly, Yu et al. (2015) uncover linguistic signals of deception in the online *Killer Game*, where teams of killers, detectives, and citizens work to convince each other of their roles over several rounds. Their subgroup detection method outperforms humans at detecting signals of deception. In another interesting online game, Cadilhac et al. (2013) predict player trades in *Settlers of Catan*, a game where players trade resources with each other to develop their regions and earn points. Their model uses dialog acts to dynamically estimate player preferences as the game unfolds. These papers are indicative of the kind of work that can be carried out within sociology to better understand social action and the dispositions that precede it.

Textual clues have also been used to identify the roles of actors within social games as well as strategies to make actors more effective in their roles. Cheng et al. (2015) use data on banned members from three different online communities to predict antisocial individuals or "trolls". They find that trolls "tend to concentrate their efforts in a small number of threads, are more likely to post irrelevantly, and are more successful at garnering responses from other users" (Cheng et al. 2015). Similarly, Blackburn and Kwak (2014) predict toxic behavior in the *League of Legends* online game by using a supervised learning approach on the decisions of over 10 million user reports involving 1.46 million toxic players. In a different context, Mukherjee et al. (2014) use a Markov Random Field model to establish which user-generated medical statements on one of the largest online health communities are credible and trustworthy. Yang et al. (2015) study teams in two Massive Open Online Courses (MOOCs) to identify latent conversational roles played by students. After discovering these roles, and inferring their optimal distribution within a team, they validated the causal efficacy of this distribution by designing more efficient teams for subsequent projects based on those role distributions. Finally, Wallace et al. (2013) study the relationship between micro-roles, or topic and speech act distributions, within established doctor-patient interactions that improve or harm antiretroviral medication adherence.

Together, this research uses text and related communicative traces as inputs to models that detect underlying regularities in the states of individuals and collectives—the social world on which social games are played. These approaches suggest unrealized opportunities to predict other states of the social world underlying communication, including shared preferences, discriminatory biases, and a wide range of cultural assumptions that filter communication.

Discussion

So much of the social world is mediated or traced by digital text today that it has come to represent a major channel through which sociologists can understand social dynamics of the present and past. Ignoring the potential for text to illuminate our sociological understanding of virtually any contemporary social domain—from culture, courtship and sexual encounters to commerce, politics and science—would be closing our eyes to the primary data stream that social media, information and big data companies use to deliver actionable insight to all sectors of the knowledge economy (Bail 2014; Golder & Macy 2014). Moreover, attempting to analyze the expanding universe of text through conventional reading stretches the limits of human capacity. This forces the qualitative analyst to sample at rates that make it difficult to reach robust, generalizable conclusions. Here we have surveyed some of the most exciting computational approaches to text analysis, as well as their application in research that seeks sociological insight.

These tools have been used in three broad ways. Most frequently, they have been used to trace collective attention and reasoning through analyzing *content* within communicated text. A burgeoning collection of studies go beyond content, using interactive text to analyze social relationships revealed through the *process* of communication. Finally, a third collection of studies reaches deeper, making inferences about social states through *signals* hidden within communicated text. These three approaches point to different levels of sociological inference. Nevertheless, much of this research has not been performed by sociologists, but rather computer scientists who hold a commitment to building new tools and demonstrating them suggestively on large, but not always well-curated samples of text and related content. For example, in Tsur et al’s analysis of differences between the framing strategies of Democrats and Republicans, they left “detailed analysis of the interplay between the different frames… for political scientists” (2015). This interpretive hand-off highlights an opportunity for sociology. By leaving careful data collection, substantive interpretation and theoretical implications to the social scientists, this work invites sociologists to engage with this community and its models.

Beyond simply testing or extending existing sociological theory, NLP and ML approaches to text analysis have the potential to generate enormous quantities of socially relevant data from a wide range of contexts. We have shown how supervised prediction models can now reliably identify power-differences, preferences and dispositions from text and related content. Unsupervised models are generating associations of increasing complexity and accuracy, like topic models and word-embedding spaces that capture stable cultural associations known to society as a whole, but not any one person. We also point to new ML opportunities for the analysis of other traces of human behavior, interaction and communication beyond text from images, audio, video, and other digital social data such as “likes” on Facebook or “swipes” on Tinder. We hope to have demonstrated that while ML and NLP cannot reproduce the subtlety of a creative researcher, who brings a life of prior associations to their analysis, computational methods trained on big data can generate many suggestive, subtle associations beyond the sensitivity of human perception and the capacity of human memory.

The wealth of new data this is making available about social relationships, cultural associations, micro-states and behaviors poses an unique new opportunity for the construction of “grounded theory” (Glaser & Strauss 1967) in virtually every substantive domain of sociological inquiry, from social, economic and political life to organizations and social movements, inequality, race and gender. This is why we titled our review “Machine Translation: Mining text for Social Theory”. Machine learning is enabling the translation of text in to social data, which is increasingly being “mined” for theoretical possibilities.

2 Language Information Density, Conceptual Space, and Communicative Speed

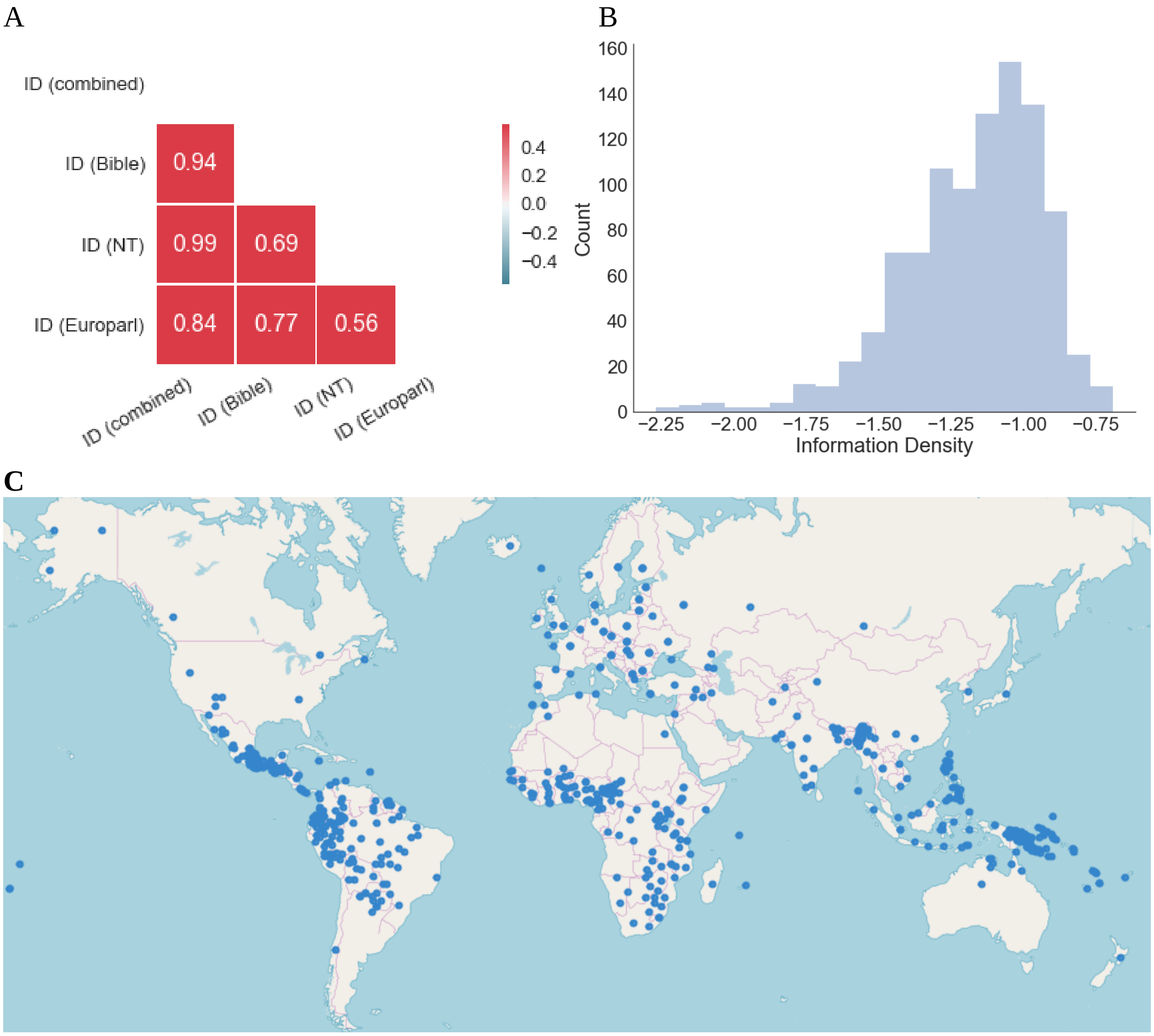
The words of a language carve meaning into the vast space of possible concepts. Much work has investigated differences in how languages encode conceptual information by comparing tightly circumscribed conceptual subdomains such as color (Davidoff et al. 1999; Kay et al. 1997; Winawer et al. 2007a), sound (Dolscheid et al. 2013), number (Bock et al. 2012), body parts (Enfield et al. 2006), human locomotion (Malt et al. 2008, 2014), time (Casasanto & Boroditsky 2008; Fuhrman et al. 2011; Lai & Boroditsky 2013), everyday activities (Majid et al. 2008; Saji et al. 2011), and space (Feist 2008; Levinson 2003). But the conceptual space encoded by a language is far larger, more multidimensional, and more nebulous than the neat categorization schemes studied within any locally circumscribed conceptual subdomain (Hofstadter & Sander 2013). Recent work has begun to address the lack of formal characterizations for the global structure of conceptual spaces (Youn et al. 2016), which are by nature multidimensional, but these attempts have remained limited to characterizing relatively small areas within conceptual space. Here I put forward one formal characterization of the structure of a language's conceptual information encoding, namely the average conceptual information density of the words within a language, and trace how variation in this structural characteristic is associated with a globally denser conceptual space and a faster rate of spoken information transfer.

Directly comparing how languages differently encode conceptual information requires the use of translated documents. The purpose of translations is to "preserve the force of [an] utterance…in forms appropriate to the target language and culture" (Bellos 2011, p. 72). To measure the average amount of conceptual information contained within the words of a language, I therefore compare how documents containing the same information content—namely the complete Bible (214 languages from 29 language families), the New Testament (828 languages from 87 language families), and the European Parliamentary Proceedings from 1996 through 2011 (21 languages from 2 language families) (Koehn 2005) —are differently encoded. These texts are ideal for this purpose because they cover large areas of conceptual space and have been translated into many languages. I estimate this measure by using the Huffman coding algorithm (Huffman 1952), an information theoretic measure that translates the word symbols of each language into the most efficient binary code possible. By reconstructing each of these documents with their Huffman codes and calculating the bit size of each document, we can say that for languages encoded in fewer bits each bit is standing in for more conceptual information, making it a denser language. To average across documents into an overall estimate of the language information density rate, for each of these measures I use English as the baseline language and convert every other language to a ratio value that uses the mean English value for that document type as the denominator (note: I use the additive inverse of the Huffman Code Density so that larger values equate to higher density). See the section "Independent Variable: Language Information Density" in chapter 3 for a mathematical description of this measure and see appendix A for a fuller description of the data sources and the text pre-processing procedure.

Figure 2.1A shows that the Huffman code information density rates are significantly correlated across documents, with Pearson's correlation r values across documents ranging from 0.56 to 0.77 (all at $p < 0.001$). Figure 2.1B displays the distribution of the information density values that have been aggregated across documents. The distribution is slightly skewed to the left, likely because of a tight upper-bound on how informationally dense a language can become.

Figure 2.1: Language information density.

(A) Pearsons's correlation coefficients across document types for the information density measure. (B) Distribution of the combined information density values. (C) Map showing the location for 686 languages (out of 986) for which latitude and longitude data was available in the WALS database (Dryer & Haspelmath 2013).

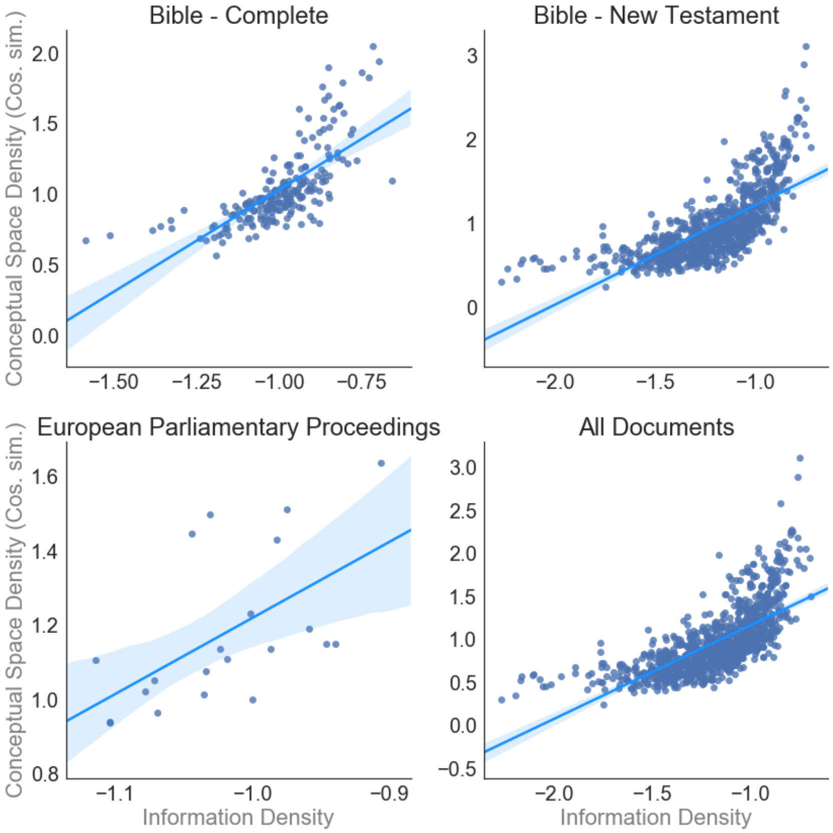


To test the relationship between information density and the density of the conceptual space, I created a 200-dimensional vector-space model of the text within each document used to construct the information density measure, using a continuous bag-of-words approach (Mikolov et al. 2013c, a,d). These kinds of models project the word co-occurrences within a text into a multi-dimensional vector space wherein similar syntactic and semantic words tend to be close to each other and wherein words can have multiple degrees of similarity (Mikolov et al. 2013a). Because these models can capture multiple degrees of similarity, they provide a useful representation of the conceptual space of a language as captured in any given corpus, given the multidimensional nature of word meaning. Vector-space models such as this effectively represent the conceptual relationships between the words of a language, capturing semantic regularities such as: V-king – V-man + V-woman = ~V-queen. The expectation is that languages with informationally denser words will tend toward conceptual spaces that are likewise denser, as each word within the space is likely to have on average more associational possibilities to every other word. The associational possibilities in higher information density languages are partly a consequence of words that appear in multiple contexts and that help to bring together disparate locations within the multidimensional spaces of these models.

Once the vector-space model for each individual document has been learned, I measured the average cosine similarity (which ranges between -1 and 1) between one thousand random word-pairs, with larger values indicating closer distances. This cosine similarity measure characterizes the average distance within the multidimensional conceptual space of each document. A larger cosine similarity indicates that concepts tend to share closer meaning associations to each other, with a more compressed space between concepts. A smaller cosine similarity, on the other hand, indicates that more space has to be travelled in order to reach any given conceptual location.

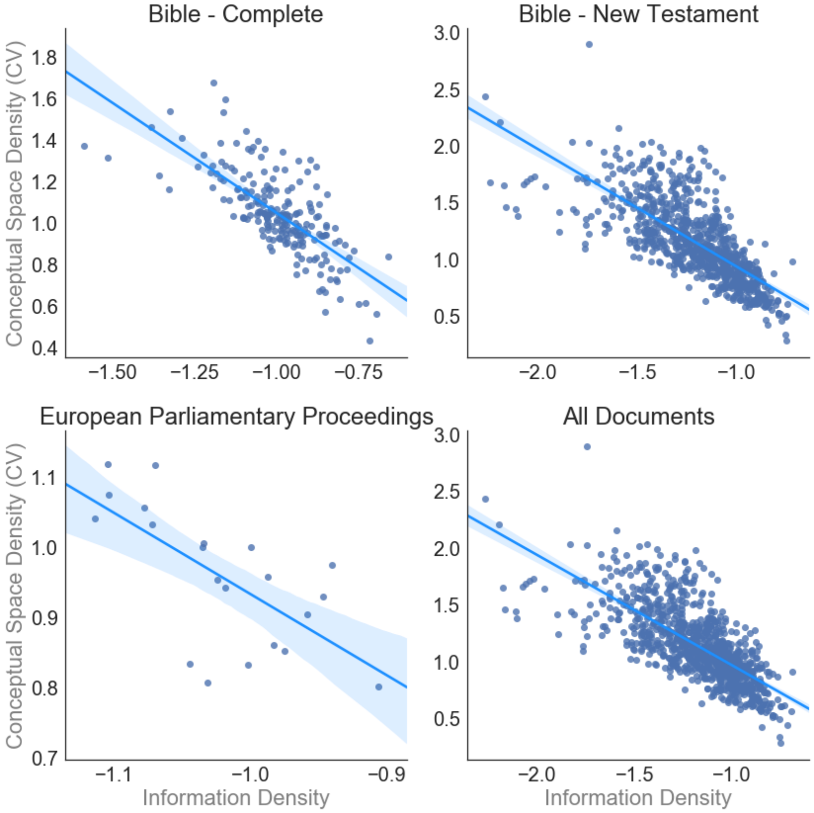
Figure 2.3 shows that across all document types, there is a strong positive association between the information density of the words of a language and the density of the conceptual space. In line with expectations, more informationally dense languages tend to have denser conceptual spaces (All Documents: r = 0.7, $p < 0.001$). To account for the possibility that cosine

Figure 2.3: Languages plotted by information density and conceptual space density (cosine similarity).



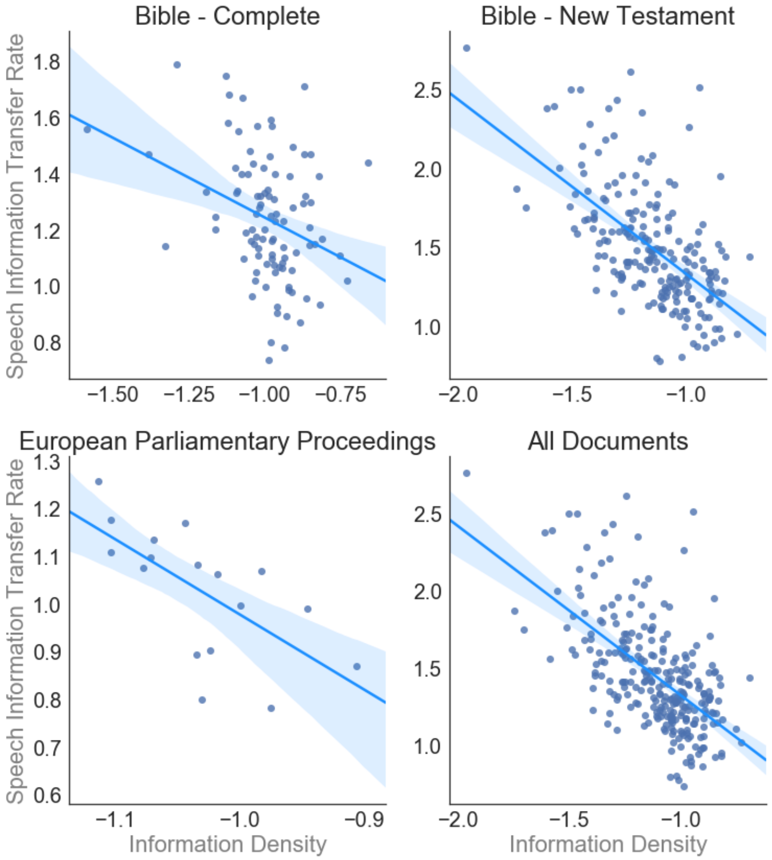
similarities are not comparable across documents, I measured the coefficient of variation for the distribution of the random word pair distances to normalize the cosine similarity parameter above. Here smaller values equate to denser distance distributions. Figure 2.4 shows consistent results across all document types (All Documents: r = -0.72, *p* < 0.001). To account for potential non-independence across language families, the above analysis was carried out by using the language family means instead of the language-level means. Results confirm the relationship (Cosine Similarity: r = 0.75, *p* < 0.001; Coefficient of Variation: r = -0.81, *p* < 0.001).

Figure 2.4: Languages plotted by information density and conceptual space density (coefficient of variation).



To test the relationship between information density and the speech information transfer rate, I collected duration times for the spoken New Testament (555 languages) and the spoken Universal Declaration of Human Rights (29 languages), providing matching information for the information density measure for a total of 272 languages stemming from 52 language families. The drawbacks of using read-aloud texts such as these include a lack of contextual variation in the text (e.g., these texts contain no humor or business language), a lack of individual speaker-level variation in paralinguistic parameters such as attitudes or emotions, and a lack of cognitive effort by the speaker to choose their own words (Pellegrino et al. 2011). Yet, these drawbacks are counterbalanced by the fact that all recordings are produced with the same intention of transmitting the same information clearly to the listener, leading to comparable data that permits for the estimation of a baseline relationship between the informational and spoken encodings of these languages. The expectation is that languages with informationally denser words will tend

Figure 2.5: Languages plotted by information density and speech information transfer rate.



toward faster speech information transfer rates due to their increased encoding efficiency. Consistent with this expectation, figure 2.5 shows that across all document types, there is a strong negative association between the information density rate and the speech information transfer rate (Bible Complete: r = -0.33, p < 0.001; New Testament: r = -0.59, p < 0.001; European Proceedings: r = -0.66, p < 0.001; All Documents Combined: r = -0.60, p < 0.001). As with the analyses above, I accounted for potential non-independence across language families by using the language family means, finding the same negative relationship (All Documents: r = -0.57, $p < 0.001$)

Across a broad sample of world languages, I show that there is significant variation in their rate of lexical information density, and that this rate of information density is positively associated with the density the conceptual space and negatively associated with the speech information transfer rate. Lexicalized concepts are of central importance to cognitive science, as these are some of the few mental representations to which we have immediate access. A prevailing assumption among those studying conceptual cognition has been that words function as a window into the study of conceptual thought (Malt et al. 2015). Yet, as others have pointed out, most people do not live in Western, Educated, Industrialized, Rich, and Democratic (WEIRD), societies (Henrich et al. 2010a,b), let alone English-speaking WEIRD societies. While prior research on conceptual variation across languages has focused local subdomains of knowledge such as color (Davidoff et al. 1999; Kay et al. 1997; Winawer et al. 2007a), sound (Dolscheid et al. 2013), or number (Bock et al. 2012), this chapter finds the existence of variation across the global structure of conceptual space, suggesting the possibility that varying domains of cognition reliant on conceptual structure—from attention to learning to memory—might in turn also vary across languages.

3 Linguistic Relativity, Collective Cognition, and Team Performance

In the lives of individuals and societies, language is a factor of greater importance than any other (de Saussure 1986, p. 7).

Groups play a central role in social, economic, and organizational activity (Fine 2012), and as a result, researchers have been interested in understanding what determines their success and failure. To date, most explanations for group performance have been investigated from one of two perspectives. On the one hand, structural approaches tend to examine the problem from the top down, studying how the position of groups within the broader environment affect outcomes. Such research has emphasized the role of structural attributes such as embeddedness (Krippner & Alvarez 2007; Uzzi 1997), brokerage (Burt 1992; Stovel & Shaw 2012), and cohesion (Burt 2005; Reagans & McEvily 2003) to account for performance differences across groups. On the other hand, compositional approaches have stressed a bottom-up perspective, tracing how group members and their characteristics influence communication processes during social interaction. Such research has traced how the distribution and diversity of member attributes and knowledge shape performance outcomes (Bell 2007; Mathieu et al. 2008). For example, the distribution of personality types (Hofmann & Jones 2005; LePine 2003), the diversity of a team's composition (Hong & Page 2004; Jehn et al. 1999; Pelled et al. 1999), and the knowledge specialization of group members (Reagans et al. 2016; Ren & Argote 2011) have been shown to affect group performance.

Although these approaches examine group performance from different perspectives, both emphasize the role that information transmission plays in enhancing group performance. Given this shared emphasis on information transmission as the driving mechanism for group performance, it is surprising that almost no attention has been paid to the role that language structure plays in shaping group processes, since language is the primary communication medium used by groups. Language, understood as “a system of signs expressing ideas” (de Saussure 1986, p. 15), is the glue that holds social life together. Through language, individuals gain access to a large set of mutually understood cues that can be used to shape and control their own mental representations and to communicate these representations to others (Lupyan & Bergen 2016). Because mental representations shape how individuals and organizations understand the world and make decisions within it (Csaszar & Levinthal 2016; Gavetti & Levinthal 2000), how these representations are encoded by language stands to shape the nature of collective cognition. By collective cognition I simply mean taking the group as the unit of analysis for cognitive activity, and applying the principal metaphor of cognitive science—that of cognition as computation—to the group’s activities by tracing how groups process information during collective tasks such as search, creativity, problem-solving, and decision-making (Hinsz et al. 1997; Hutchins 1995; Wegner 1987; Weick & Roberts 1993; Woolley et al. 2010b).

Across millennia, human groups have evolved a diverse array of language structures (e.g., sound patterns, word structures, grammatical rules (for a review, see Moravcsik 2013)) to transmit information (Campbell 2013). Work in cognitive science and linguistics describes how language structure can shape individual human thought. The long-researched linguistic relativity hypothesis predicts that the structure of a person’s language influences their cognition (Gumperz & Levinson 1996; Lakoff 2008; Lucy 1992a,b; Whorf et al. 2012). For example, Chen (2013) found that languages that grammatically associate the future and the present (these are languages where you would say “it *is raining* today” and “it *is raining* tomorrow,” rather than “it *will rain* tomorrow”), foster higher rates of future-oriented behaviors such as higher savings rates and lower smoking rates.

While the linguistic relativity hypothesis has only been pursued in the context of how language affects individual cognition, I bring the argument into social territory by moving beyond the individual to focus on patterns of collective cognition, asking: 1) Can differences in language structure affect the performance of groups? If so, 2) what collective cognition mechanisms are likely to account for this performance difference? To this end, I draw on linguistic, cognitive, and information theory to develop a novel measure of language structure, information density, to examine how language structure can shape the nature and flow of information between individuals and thereby affect group performance.

I conceptualize language information density as the average amount of conceptual information contained within the words of a language. In more informationally dense languages, words are, on average, connected to more concepts than words in less informationally dense languages. For example, if we consider the same information content, that a person plays soccer and the violin, the English word “play” is connected to the *soccer* concept as well as the *violin* concept, as one can play both soccer and the violin. However, in Spanish the word “play” (i.e., “juego”) is only connected to the *soccer* concept, with “touch” (i.e., “toco”) being the verb related to the *violin* concept, making Spanish the less informationally dense language in this example. A higher information density language, then, has the consequence of compressing the conceptual space, which is the multidimensional space within which concepts relate to each other (Hofstadter & Sander 2013). The example above shows that the same conceptual space in relation to playing soccer and the violin is more compressed in English, given that one word, “play,” stands in for more conceptual information, thereby helping to bring the *soccer* and *violin* concepts closer together. I expect that this rate of language information density will matter for processes of social interaction and collective cognition due to the associative semantic search process that underlies human conceptual cognition and is materialized when individuals engage in conversation.

Cognitive theories of associative semantic search describe how individuals recall information from memory by activating connected sequences of individual items (e.g., concepts like *red*, *heart*, or *love*), with each new item in memory expanding the window of new retrieval possibilities (Abbott et al. 2015; Hills et al. 2015). Such a process of information search can be modeled as a random walk through the transition probabilities of the items in memory (Abbott et al. 2015; Hills et al. 2015). Because high information density languages entail a more equal distribution of transition probabilities across concepts, these languages should facilitate movement through the conceptual space as groups converse. In the example above, an association between the *soccer* and *violin* concepts might be expected to occur with higher probability in English than in Spanish, given that they are connected through the word "play." This ease of movement through conceptual space may then lead to better group performance in collective cognition tasks requiring effective exploration and generation of ideas during conversation (McGrath 1984). For example, collective search benefits from broad coverage of a decision-space (Rivkin & Siggelkow 2003), collective creativity benefits from novel recombination (Simonton 1999), collective problem-solving benefits from the effective mobilization of useful sets of possible solutions (Nickerson & Zenger 2004), and collective decision-making profits from the existence of diverse information inputs (Lorenz et al. 2011). If the goals of naturally occurring, long-lived groups are dependent on efficiently engaging in these kinds of collective cognition processes, then I would expect that all else equal, groups speaking higher information density languages should exhibit better performance outcomes.

I test the role that language information density plays in shaping group performance on the outcomes of all monolingual mountaineering expeditions (n=1,626) to the Himalayas from 1907 to 2015. Mountaineering expeditions are a good test case for studying the effect of language structure on group performance for several reasons. First, this is a context where natural groups are working to solve real problems over the course of months outside of a short-lived laboratory setting. During this time, expeditions are continually making strategic decisions by searching for and evaluating courses of action under conditions where almost nothing is given or easily determined (Mintzberg et al. 1976; Schwenk 1984), and everyday must problem-solve unexpected circumstances as they arise (Boukreev & DeWalt 1999; Krakauer 1998; Viesturs & Roberts 2007). Second, all expeditions are seeking to achieve the same goals on the same mountains by performing the same kinds of tasks and solving the same kinds of problems. The clear-cut nature of the performance measures (i.e., proportion of group members that summit and speed to the summit) overcome many difficulties of cross-national research, where it is not always clear what should count as success or failure. Third, this is a context where communication is important, and therefore language structure can be expected to be a key driver of group success and failure. Finally, it is a context where one can observe variation in the independent variable of interest, given that these expeditions speak different languages.

Theory

Group Performance

Prior research has focused on how the social structure and position of groups and their members affect performance outcomes. For example, structural attributes such as embeddedness—economic action that is embedded in social relations (Granovetter 1985)—can lead to greater trust, more fine-grained information transfer, and joint problem-solving arrangements (Gulati & Sytch 2007; Uzzi 1997). Brokerage—bridging connections between actors—can enhance group performance by allowing more diverse and non-redundant information to travel through the network (Burt 2004, 2005; Stovel & Shaw 2012). Finally, a high degree of cohesion can increase the willingness to transfer knowledge across group members due to reputation concerns and the emergence of cooperative norms (Reagans & McEvily 2003). Underlying each of these structural explanations for performance is the recognition that the flow of information plays a critical role in enhancing group performance.

In addition to structural work, there is a large amount of research tracing how group composition characteristics such as personality factors (Hofmann & Jones 2005; LePine 2003), age (Jehn & Bezrukova 2004), member experience (Kilduff et al. 2000), and team norms (Goncalo et al. 2015) influence group performance outcomes (for reviews see Bell 2007; Mathieu et al. 2008; Srikanth, Harvey, and Peterson 2016). A major driving mechanism within this research stream is the way team composition influences the flow of information within a group. For example, diverse groups can lead to favorable group outcomes because they have more diverse information that can be brought to bear on a problem (Hong & Page 2004; Jehn et al. 1999; Pelled et al. 1999). However, diverse groups can also inhibit performance by reducing the unity of the group and reducing communication and information transfer between members (Ancona & Caldwell 1992).

Both structural and compositional explanations of group performance posit the flow of information as a primary mechanism for the effects they catalog because it affects the transmission, interpretation, and mobilization of information within the group and between the group and the environment. In this respect, language, as the primary communication medium used to transmit information, might be expected to play a central role in group dynamics and performance. Yet, almost no attention has been paid to the role that language structure and language use play in shaping collective cognition processes and performance outcomes of groups. I here propose that the structure of language can indeed be deeply consequential to social interaction and communication processes, and to ensuing group performance outcomes. This expectation follows from research within cognitive science and linguistics that has traced how language structure can influence individual human cognition.

From the Linguistic Relativity of Individual to Collective Cognition

Human groups have evolved a diverse array of language patterns, including patterns of prosody, morphology, and syntax to transmit information (Campbell 2013). There is a long-standing interest dating back to the early 20th century concerned with tracing the effects differences in language structure have on cognition (Gumperz & Levinson 1996; Lakoff 2008; Lucy 1992a,b; Whorf et al. 2012). The linguistic relativity (or Whorfian) hypothesis asks whether structural morphosyntactic differences between languages affect habits of thought (Lucy 1992a) such as perception (Levinson 1996; Thierry et al. 2009; Winawer et al. 2007b), classification (Brown & Lenneberg 1954; Lucy 1992b), spatial orientation (Levinson 1996), and judgment (Casasanto & Boroditsky 2008).

The linguistic relativity literature has traced two primary mechanisms linking language and thought (Lucy 1997). On the one hand, the structure of language shapes how we interpret the reality experienced by the senses. The structure of language here functions as a lens, refracting information from the environment in patterned ways depending on the language being used. For example, Russian forces a distinction between lighter and darker blues ("goluboy" versus "siniy"), and therefore Russian speakers are faster to distinguish between two colors when they fall into different linguistic categories than when they do not. English speakers, who do not have such a language distinction, exhibit no such effect (Winawer et al. 2007b). On the other hand, the structure of language shapes the thoughts that one can have about reality. Here, a more appropriate metaphor is that of a train running on tracks, where the track system of language helps to determine the places that our thoughts can reach and how those places are reached (Enfield 2015). For example, research shows that speakers speaking languages that refer to spatial locations through fixed coordinates such as "to your west" rather than through relative language such as "to your left" are better at pointing in the direction of a familiar location such as the home village when located in an unfamiliar location far from home (Levinson 1996, 2003).

These mechanisms are concerned with how language structure can affect individual cognition. It is conceivable, however, that language structure is also influential during social interaction by shaping the communicative constraints and possibilities available to groups speaking different languages. In this article, I begin to address this possibility by introducing a new language structure attribute that I expect consequentially affects the communicative possibilities available during social interaction. I theorize how the information density of a language shapes how groups navigate the conceptual space (Hofstadter & Sander 2013) during conversation and the effects that this movement through the conceptual space is expected to have on the group's collective cognition and performance. The conceptual space of a language, which is more deeply described below, is "the multidimensional space within which concepts exist"

(Hofstadter & Sander 2013, p. 50). The semantic meaning of concepts exists as a function of how concepts relate to each other within this multidimensional space. For example, if English were conceived as a 300-dimensional space (Mikolov et al. 2013b), one could take the vector for "king," subtract the vector for "man," add the vector for "woman," and arrive closest to the location of "queen." It is through the words of a language that this conceptual information is communicated and computed upon by groups as they collectively think during conversation. Below I argue that how they do this will depend on the information density of the group's language.

Language Information Density

Differences in the semantic meaning of a word can be subtle and often unnoticeable, such as in the sentence below (borrowed with slight modifications from Hofstadter and Sander 2013, p. 10):

1) I play soccer, Monopoly, and violin.

In this sentence, one could argue that the word "play" has two distinct meanings, one in relation to soccer and Monopoly (given that they are both games), and the other in relation to violin (which is not a game). While English does not specify this difference, Spanish does. In Spanish the verb "juego" is used for soccer and Monopoly while the verb "toco" is used for violin, giving the following sentence:

2) Yo juego soccer y Monopoly, y toco el violín.

If one wanted to create a language that specifies the *play* concept further, different verbs could be used for soccer and for Monopoly, one referring to board games and another referring to sports. Such a step would then entail three differentiated verbs rather than the one (i.e., “play”) that we started with.

As this example illustrates, the same information is being communicated in both languages even if it is being communicated through different linguistic codes. That is, the Spanish listener receives *the same* message as the English listener, even if the encoding of the message is different. Because the information content remains constant, we can say that at the level of words, the English message is encoded more efficiently given that it is encoded in six words and the Spanish message is encoded less efficiently, as it is encoded in nine words. In this example, even if we removed functional words such as “I,” “and,” “Yo,” “y,” and “el,” English would still be encoded more efficiently with four words compared to the five words necessary in Spanish[20]. I thus take the encoding efficiency of a language to be a measure of the information density that exists within that language. More formally, the information density rate of a language is the average amount of conceptual information contained within the words of a language. What this example shows is how languages can have differing degrees of information density within their lexical distributions. Below I will describe how I use mathematical tools from information theory to estimate the information density rate for each of the world’s languages.

If the sentences above were the entirety of the language, English would be said to have a higher language information density rate and hence a more compressed conceptual space, since each lexical item would represent greater amounts of conceptual information. That is, the word “play” is standing in for the concept having to do with kicking a ball (i.e., the “juego” concept) as well as for the concept of moving a bow across the strings of a violin (i.e., the “toco”

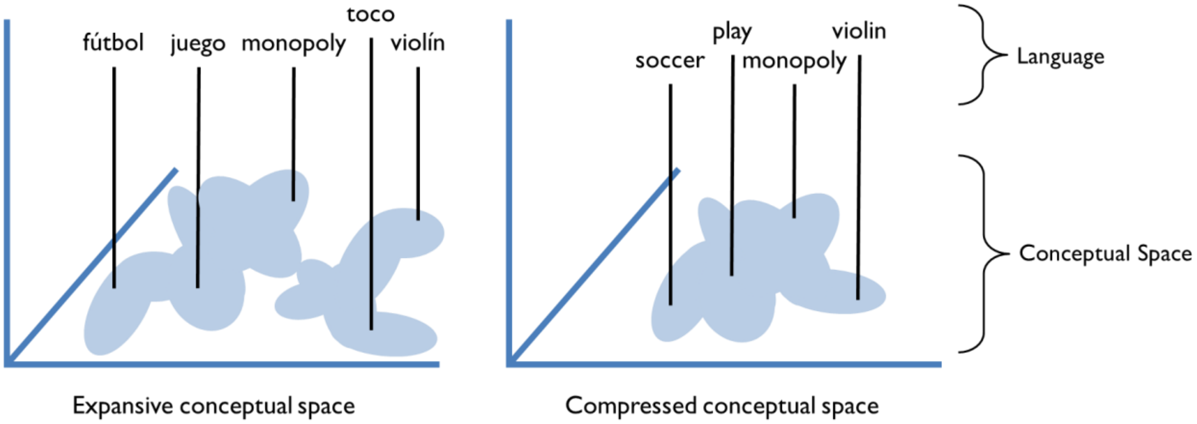
[20] I use this example to illustrate the nature of what the measure is capturing, not how I in fact capture the measure, which is fully described in the methods section below.

concept). Figure 3.1 graphically illustrates the distinction between a sparse and a dense language and the effect that the encoding has on the corresponding conceptual space. If we imagine a flat, square surface as the conceptual space that all languages encode, then different languages will fill in that space with blobs of different shapes and sizes[21]. For example, the English word “play” is connected to the *soccer* concept as well as the *violin* concept. However, in Spanish the word “play” (i.e., “juego”) is only connected to the *soccer* concept, with “touch” (i.e., “toco”) being the verb related to the *violin* concept. Reaching the *monopoly* location on the conceptual space when the conversation started at the *violin* location should be easier to do in English (the more compressed conceptual space) than in Spanish. I theorize that the compression of the conceptual space conditions the communicative possibilities of groups speaking different languages. The expectation is that languages with higher information density will facilitate group movement through the conceptual space during tasks that require effective exploration of conceptual space during conversation[22].

Figure 3.1: Example of sparse and dense information density languages and the corresponding structure of the conceptual space.

[21] For an extended discussion of this visualization of conceptual space, see page 78 in Hofstadter and Sander (2013).

[22] Some may argue for the possibility that higher information density languages may in fact increase rates of confusion between group members. I find this expectation unlikely for the following reason. During the process of human communication, speech encoding (i.e. articulation, or the process of turning information into speech sound) is the slowest part of speech production and comprehension (Levinson 2000). Pre-articulation, parsing, and comprehension run at much faster speeds. This means that humans can think of what to say and understand what somebody else has said approximately four times faster than it takes to say it (Levinson 2000, p. 6). Additionally, humans are effective meaning-disambiguators, restricting domains of subsequent reference as they hear an utterance (Altmann & Kamide 1999). This all suggests it is unlikely that confusion would be so widespread as to influence team performance, especially for tasks requiring effective creativity and problem-solving.



Language Information Density and Group Performance

In many contexts, teams and organizations depend on the efficient exploration of conceptual space to achieve optimal performance outcomes (Lavie et al. 2010; March 1991). This is especially likely to be the case when group performance is dependent on efficient search, creativity, problem-solving, and decision-making, as these collective tasks benefit from diverse information inputs, the broad coverage of a solution-space, and novel recombination of information. In this light, I argue that one way through which the information density of a language is likely to influence group performance is by affecting how a group moves through conceptual space during conversation. When groups are engaged in cognitive tasks that require efficient exploration, I expect that groups speaking higher information density languages will be able to more easily traverse through conceptual space during conversation, thereby leading to broader and more efficient exploration of the space. But what reason have I to expect that these processes of collective cognition will vary with the information density rate of the group's language? The answer lies in the role that words play during the communicative process, in the cognitive underpinnings of long-term memory retrieval, and in the way language information density shapes the conceptual space. I address each of these issues in turn.

At the neural level, there is no such thing as a pre-existing conceptual space that all humans share (Lupyan 2012), as common biology and environment underdetermine the possible concepts with which humans represent the world (Malt et al. 2015). Rather, the words of a language are responsible for carving up the conceptual space (Lupyan & Bergen 2016). This means that the words determine which parts of the conceptual space will be represented and how the conceptual space itself is to be partitioned. From the color example above, we saw that the location of conceptual space where the concept *blue* is located is differently carved in Russian and English, with the Russian region of conceptual space being represented by more words representing smaller areas of the space (e.g., “goluboy” and “siniy”) compared to the larger and more encompassing English word “blue.” Humans, then, use words to refer to concepts, which are dynamic patterns of information that are made active in memory in response to internal or external cues (Casasanto & Lupyan 2015). Words gain their significance by being connected to concepts, and are inputs to a dynamic cognitive system that is ever-transitioning from one mental state to the next (Lupyan & Bergen 2016), often as a consequence of the associations that words elicit (Hills et al. 2015). Because these transitions in mental states are driven by the neural underpinnings of words and concepts, then the way words cue associations between concepts during communication (Hills et al. 2015; Hofstadter & Sander 2013) is likely to drive how groups move through a conceptual space during conversation.

The clearest evidence for how individuals move through conceptual space at the cognitive level comes from research on long-term memory retrieval (Davelaar & Raaijmakers

2012). Such research, often carried out through verbal fluency tasks[23], finds that sets of semantically similar words are generated together (Howard et al. 2007; Romney et al. 1993) through a random-walk, associative retrieval process (Abbott et al. 2015; Hills et al. 2015). This means that words function as cues to activate associations in memory. Each word or subsequent set of words activates a new set of potential words from which to proceed. Consistent with optimal foraging theory (Charnov 1976; Pirolli 2007), the process of associative semantic search is carried out by exploiting local word patches until cues are depleted (Davelaar & Raaijmakers 2012; Harbison et al. 2009) and the individual undertakes global exploration in search of new word patches. The process that occurs during conversation is an extension of this process, but instead of internal word cues, the cues come from the utterances of others. As each new word is spoken, it serves as a cue that activates new traces in the memory of all participants, opening new possibilities for movement through the conceptual space during the conversation. These possibilities will necessarily be a function of the information density rate of the language being spoken.

A fruitful way to think about the nature of the possible routes through conceptual space is through the idea of the adjacent possible (Kauffman 2000). The concept of the adjacent possible comprises "all those things, ideas, linguistic structures, concepts, molecules, genomes, technological artifacts etc., that are one step away from what actually exists, and hence can arise from incremental modifications and/or recombination of existing material" (Loreto et al. 2016, p. 60). In a real sense, the adjacent possible exists solely as a function of the actual, and "is a kind of shadow future, hovering on the edges of the present state of things, a map of all the ways in which the present can reinvent itself" (Johnson 2010, p. 31). In the context of collective

[23] In verbal fluency tasks, individuals have to produce as many words as possible from a category (e.g. animals or fruits) during a given time period, typically ranging from one to three minutes.

cognition processes that take place through conversation, the actual is composed of all the concepts that have been activated by words during a conversation and that have been registered into the collective memory of a group's conversation. The adjacent possible, then, is the space of future conceptual states that are one step away from the actual and toward which the conversation can proceed. As a conversation progresses, the adjacent possible grows larger due to the increased number of words and concepts activated and registered into the groups collective memory. As discussed above, higher information density languages contain more conceptual information within each word. Mathematically speaking, the consequence of a higher information density rate is a larger adjacent possible during conversation.

The theoretical logic established thus far involves the following. Individuals think through processes of associative semantic search, consistent with a random walk model. Languages encode conceptual information differently, with higher information density languages exhibiting a more compressed conceptual space and a larger adjacent possible. When individuals converse, words function as cues that activate the conceptual space in memory and shape the possible directions a conversation can move in. The denser the conceptual space and the larger its adjacent possible, the easier it will be to explore the space for at least three reasons. First, conversations will be more fluid due to closer distances and less friction between the concepts within the conceptual space. From our example above, a group speaking English would be more likely to reach the *violin* location on the space when the starting location of a conversation was *soccer*, given that these two concepts are connected through the word "play." Second, words will be more likely to spark distant associations. For example, in English the word "bank" activates both the concept *financial institution* as well as the concept *side of a river*. Given a circumscribed set of solutions within conceptual space, this should lead to the beneficial locations on the space being more efficiently reached in higher information density languages. Third, conversations will be less likely to get stuck. Because of the extra conceptual information within each word, after each conversational turn there are more places to which the group can move. From our example, we can see that a conversation in English that starts with the *play* concept can with high likelihood reach the *soccer*, *monopoly*, or *violin* concepts. However, a conversation in Spanish would take longer and ultimately be less likely to reach the *violin* concept from the same starting point. Since tasks such as search, creativity, problem-solving, and decision-making are ones that benefit from the exploration of conceptual space, we can therefore expect that

> ***H1****: Groups speaking informationally denser languages are more likely to succeed at tasks requiring high levels of search, creativity, problem-solving, and decision-making.*

For naturally occurring groups that engage in tasks requiring effective exploration for decision-making, it is rarely sufficient to explore conceptual space for a solution. Rather, the process tends to iterate between an identification phase where problems or opportunities are recognized, a development phase where solutions are developed or opportunities are elaborated, and a selection phase where alternatives are more deeply investigated (Mintzberg et al. 1976). Further, solutions that are eventually selected and acted upon themselves lead to a feedback process where the solution, if not fully implemented, is iterated upon and adjusted accordingly or discarded altogether in place of a renewed development and selection process. This suggests that the more efficiently a group can move through conceptual space, the more quickly it will be able to undertake each of these decision-making phases, which should lead groups speaking higher information density languages to achieve performance outcomes more quickly. It is therefore not only likely that locations within the conceptual space that have higher payoffs will be reached more frequently by groups speaking higher information density languages, but given the faster rate of iteration between the phases of the decision-making process that is possible, it is also likely that they will do so more quickly. Therefore,

> ***H2**: Conditional on having succeeded at the task, groups speaking informationally denser languages will achieve success more quickly.*

The communication process is central to the proposed effects outlined above. Scholars have long recognized that team-member characteristics can shape patterns of social interaction within a group (Berger et al. 1972, 1980; Bunderson 2003). The distribution of power, prestige, and expertise across team members can drive the performance expectations that members have for each other, shaping their understanding of each member's ability to contribute to the team's task (Berger et al. 1972). Therefore, leadership status within a group has been associated with higher rates of verbal participation within small groups, as leaders tend to have skills and knowledge relevant to the task in addition to the status and prestige attached to their social position (Stein & Heller 1979). Consequently, other team members may perceive their views and ideas to be of little value, leading to decreased communication and information sharing within the team (Tost et al. 2012). Even in groups without formal positions of authority or expertise, group members tend to give more weight to the ideas of the highest performing members, regardless of their true expertise (Bonner et al. 2002).

We might therefore expect that team-level attributes that interfere with the open transfer of information within a team— such as the distribution of status and expertise between team members—will moderate the effect of language information density on team performance. In general, teams with highly experienced leaders relative to the experience of the remainder of the team will likely experience fewer benefits from language information density due to the decreased role played by active communication within these teams. On the other hand, teams with flatter status hierarchies (i.e., teams with a smaller discrepancy between the expertise of the leader and the mean expertise of the other team-members) will tend to more actively converse with each other during consequential tasks, amplifying the effect of language information density on group performance. Therefore,

> ***H3****: The positive effect of language information density on team performance will be amplified for teams that have a flatter status hierarchy.*

Data and Methodology

Empirical Context: Mountaineering Expeditions to the Himalayas

I test the hypotheses with data from the Himalayan Database, a catalogue with the details of all expeditions and individuals that have climbed Himalayan mountain peaks from 1905 to 2015 (Hawley & Salisbury 2004). The information is composed of the archives of Elizabeth Hawley, a journalist and long-time resident of Kathmandu, who became a local institution by dedicating her life to collecting the details from all expeditions to the Himalayas. Her collected archives were digitized by Richard Salisbury, a climber and database manager at the University of Michigan, and have been supplemented with information from books, alpine records, and communications with individual climbers. As of 2015, the database included information on over 8,700 expeditions, 65,000 climbers, and 450 mountain peaks.

Expeditions to the Himalayas are a good context for examining the effects of language information density because the activity of summiting peaks requires active and ongoing communication and strategic decision-making at the team level. Efficient movement through conceptual space is crucial for summiting as strategic decision-making and daily problem-solving (Boukreev & DeWalt 1999; Krakauer 1998; Viesturs & Roberts 2007) are critical tasks that teams regularly engage in. A survey of experienced mountain climbers from 27 countries confirms this expectation (Anicich et al. 2015). When asked about the importance of team processes such as communication and exchanging information on a 7-point Likert scale, the mean response was 6.5. There are multiple stages where communication becomes critical. During the planning stage, teams must organize gear and supplies, plan logistics for the expedition, and foresee contingencies and problems that may arise to effectively prepare. During the acclimating stage, expeditions must plan their navigation and route selection, strategize climbing speed and pacing, take physiological and weather assessments into consideration, and think about and plan for as many contingencies as possible. Last, during the climbing and descent stages, teams must understand problems as they emerge, generate sets of solutions for these problems, evaluate these solutions and their likely outcomes, and decide which solution is likely to be most successful. During an expedition’s duration, which typically lasts for many months from ideation to team-building to planning and to execution, the expectation is that the effects of higher language information density on the groups collective cognition will aggregate into better expedition-level performance outcomes. Thus, during each of these stages, the better the team moves through the conceptual space, the more likely they will be to reach the summit (H1), and conditional on reaching it, to reach it more quickly (H2). Further, this effect should be amplified within low-hierarchy expeditions, given that it is here that communication is most likely to play a role in shaping team performance (H3).

These mountaineering expeditions are a good test case for the effect of language information density on group performance for a few additional reasons. First, finding contexts where language is likely to be a key driver, where groups are the primary unit of analysis, where the groups vary in terms of the information density of their languages, and where detailed data is available is difficult. In most cases, at least one of these components will be missing, yet this context has all four. Second, this context offers clear-cut, uniform performance measures including the percent of team-members that summit and the speed to the summit. Third, all expeditions are climbing the same mountains, holding the nature of the real-world task constant. Further, these expeditions are natural, long-lived groups engaging in a broad set of activities taking place over the course of months or years (McGrath 1984). This suggests a higher degree of ecological validity, at least in comparison to short-lived groups performing circumscribed and artificial tasks in a laboratory setting. Finally, the validity and usefulness of the data for examining cross-country effects has been established by other scholars, most notably Anicich et al. (2015), who use it to explore how national culture attributes affect the performance of teams.

In this chapter, I trace the performance of all monolingual expeditions to the Himalayas from 1907 to 2015. I use expeditions where all members came from the same country as a proxy for language. I exclude cases where several languages are spoken within a country (e.g., India and Canada), because the team's language cannot be inferred by country membership. In addition, I exclude all expeditions with hired Sherpas, as the process of translation would likely affect the communication process and interfere with the posited language effect[24]. I also removed

[24] Out of all expeditions in the database, 55 percent do not use Sherpas.

all expeditions to the Ama Dablam peak, which is infamous for overcrowding (Hawley 2014). Reading through the seasonal summaries written by Elizabeth Hawley (2014), which documented the major events of each climbing season from 1985 to 2014, it is clear that Ama Dablam is an outlier. Since the late 1980's, the constant over-crowding on Ama Dablam means that expeditions to this peak are not able to move at their own pace. Rather, fixed lines are set-up by Sherpas at the start of a season, and expeditions move up and down the mountain following each other at the same rate. If an expedition were to speed up, they would have nowhere to go once they reached the bottleneck on the fixed lines. Furthermore, there is limited space in the midway camps, so even if an expedition managed to move past the bottleneck, they would have no place to establish a mid-way camp. Appendix B1 includes several quotes from Hawley's seasonal documenting this perpetual overcrowding issue on Ama Dablam[25].

The expeditions in the sample (n = 1,626) came from 31 countries, representing a total of 21 languages. The mean year for these expeditions was 1997, with a standard deviation of 13 years, and with 75 percent of the expeditions occurring after 1989. From this dataset, I created many controls related to the characteristics of each expedition. I then merged this information with country-level controls for each expedition related to the national economy, characteristics of the population, and national-level cultural attributes.

Independent Variable: Language Information Density

While the English and Spanish example above is straightforward, it is impossible to establish how much conceptual information exists per word within a language, given that most

[25] The results below are unaffected by keeping Ama Dablam in the sample. Results available upon request.

word meanings are not explicitly captured in dictionaries and thesauri due to their semantic subtlety. For example, in the sentence “the book was clothbound but unfortunately out of print,” the word “book” includes two conceptual meanings “object made of printed sheets of paper,” and “the set of all copies available in stores or warehouses,” (Hofstadter & Sander 2013). One could easily imagine many other conceptual meanings for the word “book.” This is akin to the situation physicists find themselves in when examining a thermodynamic system. While they are unable to establish the precise speed and location of all the particles within the system, they are nonetheless able to use probability theory to measure the average behavior of the particles within the system, giving us a measure of temperature. Similarly, we cannot establish the precise amount of conceptual information for each word within a language, but we can estimate the average amount across all words given a large enough text sample. In this way, the information density rate of a language is like the temperature of a thermodynamic system, and just as we might compare the temperatures of different thermodynamic systems, we can compare the density rates of different languages.

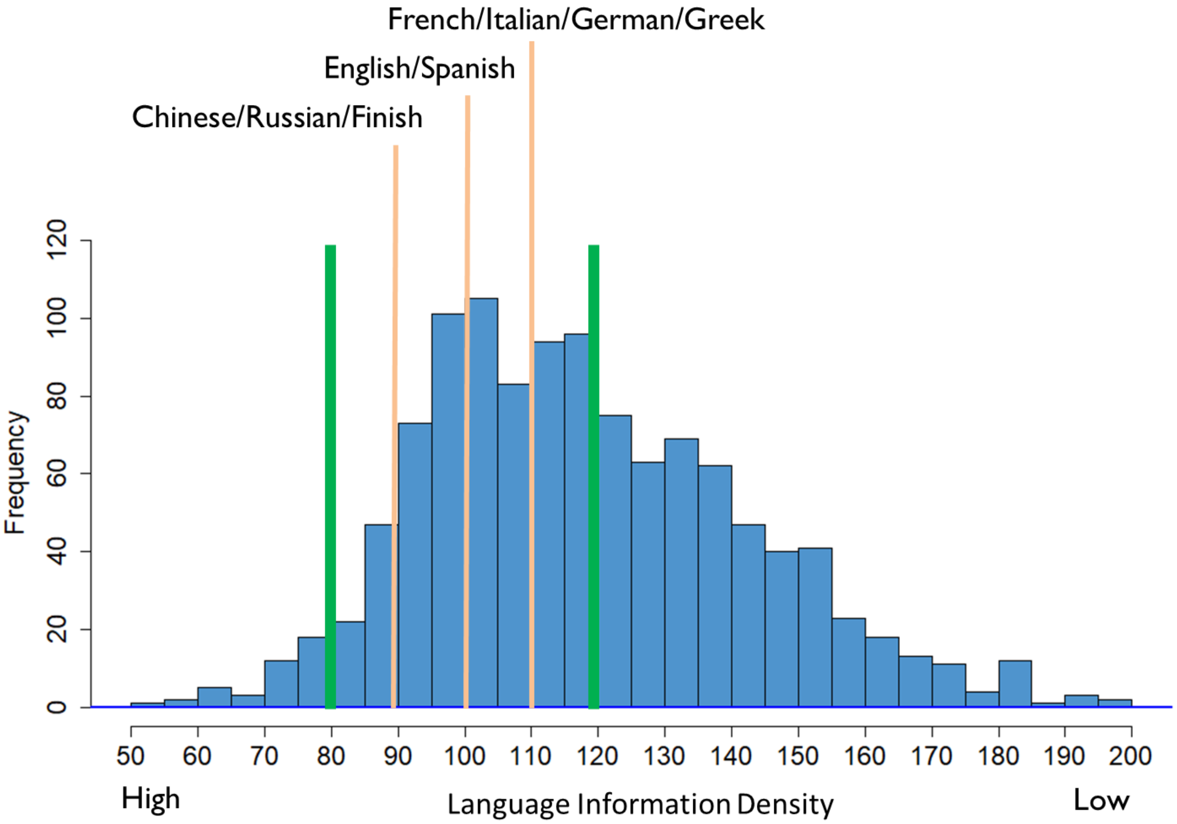
To estimate the language information density rate, I use ideas from information theory, which features established methods for characterizing the statistical nature of word distributions (Shannon 1948). To be able to measure the information density of a language, we need to be able to compare how the *same* conceptual information is differently encoded through the vocabulary of a language. Holding the information that is communicated constant allows us to trace how different languages encode that information. The development of this measure included several steps. First, to vary language structure but maintain consistent information content (i.e., the conceptual space) I sought out documents that have been translated into many languages, given that the process of translation is an attempt to communicate the same information, force, or meaning of an utterance through the use of a different code (Bellos 2011). To build the measure I use five documents: 1) the Bible, 2) a large parallel corpus of United Nations proceedings and documents, 3) and a large parallel corpus of European Parliamentary proceedings. Because each of these documents contains the same information (as this is the point of translation), then we can see how it is that each language breaks up the same conceptual space differently given its word distribution.

Second, to build the word distribution of each document, I tokenized (i.e., broke up each text into a list of words) all the words and counted how many times each word appeared in the document. I then used this distribution to generate the most efficient binary code for each document using the Huffman coding algorithm (Huffman 1952). This algorithm creates the most efficient prefix-free binary code with which to encode a given symbol distribution. Formally, the input into the algorithm is a symbol set $A = \{a_1, a_2, \ldots, n\}$ of size n, with weights $W = \{w_1, w_2, \ldots, n\}$, where $w_i = weight(a_i), 1 \leq i \leq n$. In the case of language, the symbols can be characters or words. Because my interest lies in how the conceptual space is encoded, I use the symbol distribution of words within a language. The output of the algorithm is a code $C(W) = \{c_1, c_2, \ldots, c_n\}$, where $c_i$is the codeword for a_i. We let $L\big(C(W)\big) = \sum_{i=1}^{n} w_i \times length(c_i)$ be the weighted path length of code C, and satisfy the condition $L\big(C(W)\big) \leq L(T(W))$ for any code $T(W)$. The output of the Huffman Coding algorithm, then, is the shortest possible binary code for a given symbol distribution.

When every word in a text is translated into its Huffman code, we are left with a document of 0’s and 1’s and can count the number of binary digits (bits) in each Huffman coded document. For a given document (e.g., the European Parliamentary Proceedings), the larger the bit size of a document, the less informationally dense a language is, as this means that each bit contains less conceptual information. The smaller the bit size, the more informationally dense a language is, as this means that more conceptual information is contained within each bit[26]. To aggregate the rates across documents, I take English as the baseline bit size for each document type and every other language takes on a ratio value relative to English. That is, the English value is the denominator and the ratio of each language is arrived at by using its document bit size in the numerator. I then average rates across document types into an aggregate information density rate for each language. The advantage of averaging across document types is that we get a better estimate for a language, given that different contexts of speech (e.g., historical, political) may have slightly different probability distributions across their words. I created this measure for 986 languages for which parallel translation corpora existed. For ease of interpretation of the statistical tests, the measure used in this chapter takes this ratio and inverts the sign so that higher information density is a larger value, and then normalizes the values to fall between 0 and 100. Figure 3.2 below displays the non-inverted distribution of languages across the information density continuum (multiplied by 100). The world's thirty most widely spoken languages fall within the green bars.

[26] While it does not matter for the logic of the measure and theoretical logic of the paper what structural linguistic differences engender the variation in the measure, two common ways in which more information can be embedded into each bit are ambiguity and agglutination. As the book example above demonstrates, ambiguity refers to the same word having multiple meanings. Agglutination refers to languages that use multiple morphemes within words and for which the morphemes themselves remain unchanged. In each of these cases, more information is activated for associative potential. In the case of ambiguity this happens by each word referring to multiple concepts, and in the case of agglutination it happens by each word specifying more information than the context requires. For example, the German word "Schifffahrtskapitänkabinenschlüssel" refers to the concept *key to the cabin of the captain of a ship*. If such a concept were to be communicated in English, one might perhaps only say "the key to the cabin" and let the context fill in the remaining meaning.

Figure 3.2: Distribution of languages by their information density rate (not normalized).



Independent Variable: Team Hierarchy

To test hypothesis 3, I created a measure that captures the difference between the experience level of the team leader and the experience level of the other team members. I first created a measure of leadership experience, which is the mean number of expeditions the leaders of an expedition have led in the past (some expeditions had more than one leader). I then created a measure of team member experience, which is the mean number of prior expeditions team members have undertaken. I next subtracted the mean member experience measure from the mean leader experience measure and discretized this distribution into three groups to facilitate interpretation.

Dependent Variables

The first outcome of interest is expedition success, which was measured as the count of group members from an expedition who reached the summit. The second outcome of interest is the speed, in days, that it took an expedition to reach the summit from a peak's basecamp.[27]

Control Variables

The selection of control variables closely follows the approach and controls in Anicich et al. (2015), who investigated the effect of hierarchical cultural values on Himalayan mountaineering expeditions. All controls for the environmental factors, risk preferences, and expedition attributes were coded from data in the Himalayan database (Hawley & Salisbury 2004), while controls relating to the home country of the expedition came from a variety of sources described below. First, three controls were used to account for the environmental conditions of each expedition. *Himalayan region* is included as a fixed effect to account for different levels of infrastructure in each of 20 regions where the peaks are located. Fixed effects for the *season* were also used to account for different climatic conditions. I also included a

[27] Some might argue that deaths on these expeditions might also be a meaningful measure of success. However, descriptive analysis shows that across all expeditions to the Himalayas, over 70 percent of deaths were caused by unexpected accidents such as falls, crevasses, icefall collapse, avalanches, and falling rock. As these are random events, there is no reason to think that the language information density rate should affect the death rate of expeditions to the Himalayas.

continuous measure for *year* to reflect improvements over time in climbing gear and the changing nature of climbing in the Himalayas.

Second, two controls were used to account for the risk preferences of expeditions. I controlled for whether an expedition took a *standard climbing route,* as non-standard routes can be riskier, more treacherous, and have less external support. Similarly, I categorized whether a route was *illegal* or not.

Third, many expedition attributes were accounted for. I counted the *number of climbers on an expedition*, as communication and coordination costs can increase within larger groups. The *mean age of climbers* on an expedition was used to control for fitness levels and life experience that could affect the outcomes. *Team leader experience* was operationalized by counting how many expeditions a leader has led, as leaders with more experience will have more knowledge from which to draw. Similarly, the *average experience of the expedition members* was generated by counting how many expeditions a member has been on. In this vein, I controlled for the *standard deviation of group member experience,* since large differences in experience could affect communication and coordination within the team. I also controlled for the *percent of climbers on an expedition that used oxygen*, since oxygen use affects high-altitude physiology and is likely to influence outcomes. Number of *camps* set up during an expedition was used because they allow shelter and physical and mental recovery. I controlled for the *percent of female climbers* on an expedition, because gender could be associated with varying amounts of physical strength and different styles of communication (Woolley et al. 2010b). The *height of the peak* (logged) was used to control for increased health risks associated with altitude.

Fourth, I controlled for many socio-economic and cultural attributes relating to the home country of the expeditions, which may result in advantages that contribute to team success.

Controls for home country attributes came from many sources, described in turn. *Gross domestic product per capita* (World Bank 2016) was used as it is possible that expeditions from wealthier countries have better equipment and training (while research on wealth normally logs a measure such as this due to a skewed distribution, it was normally distributed in my sample). I included a measure for the *size of the population* (logged) (World Bank 2016), as more populous countries have a larger pool of talent from which skilled climbers can be drawn. I used a measure of the *climatic demands* of a country, which captures deviations in the extent of cold and hot temperatures within a country, as it has been shown to be associated with different kinds of freedoms (Van de Vliert 2013) and could influence patterns of communication. A country's Gini coefficient was used as a measure of *income inequality* (World Bank 2016), as it could affect patterns of communication within the team, with more equitable countries communicating more openly. I also controlled for the *educational level* of a country (United Nations 2016) through an index of educational attainment created by the United Nations.

Finally, I control for several country-level cultural attributes, since cultural differences are an important driver of individual and collective behavior (Kitayama & Uskul 2011; Markus & Kitayama 1991). I control for cultural hierarchy with the *hierarchy index* measure used by Anicich et al. (2015), who show that expeditions from hierarchical cultures are associated with a higher number of members reaching the summit. In line with Anicich et al., (2015) I also control for the cultural attributes of mastery, harmony, and embeddedness put forward in Schwarz and Ischebeck (2003) and individualism, masculinity, and uncertainty avoidance put forward in Hofstede (1984, 2001). Each of these cultural attributes could be posited to affect patterns of communication within the teams.

Analytical Approach

The first dependent variable is the count of expedition members who reached the summit. As overdispersion is present in this outcome, I first compared whether a poisson or negative binomial model would be more appropriate to model the association between language information density and the number of group members who reached the summit. In this case, a negative binomial model is more appropriate, as the likelihood ratio test that alpha equals zero was rejected ($p < .0001$). Further, approximately 58 percent of observations for this outcome had a value of zero. I therefore tested to see whether it would be appropriate to use a zero-inflated count model, which responds to the failure of the negative binomial model to account for excess zeros by changing the mean structure to allow zeros to be generated by two distinct processes (Lambert 1992). Such an approach makes sense in this context, where many expeditions are terminated because of bad weather such as storms and high winds, or because of bad conditions such as deep snow and avalanching. It is therefore appropriate to assume that there are two latent groups in the sample, one for which an outcome of zero had a probability of 1, and one for which the outcome is zero but nonetheless had a non-zero probability of having a positive outcome. Results from the Vuong test (Vuong 1989) support the use of a zero-inflated negative binomial modeling strategy (V = 4.58, p = .0000). Accordingly, to examine H1 I use a zero-inflated negative binomial model (ZINB) with expedition size to account for inflation and clustered robust standard errors to account for the fact that expeditions are nested within countries (Anicich et al. 2015; Rogers 1994).

The second dependent variable is speed to the summit. Of the 1,626 expeditions in my sample, 628 reached a summit. These expeditions took place on 145 separate peaks, and of these

133 had fewer than five expeditions. Because each peak has a different distribution of the outcome variable and because of the small number of observations on some peaks, I compare within-peak variation on the two most widely climbed peaks, Cho Oyu (n = 160) and Ama Dablam (n = 155). As described above, Ama Dablam is a unique mountain to climb because there is little that teams can do in terms of their communication to reach the top with greater likelihood or to get to the top more quickly relative to any other team. The route to the peak has fixed lines that are set at the beginning of the season and to which all expeditions have equal access. Further, there is limited room on the route and on the camps on the way up and down, which means that expeditions essentially move along this corridor as if on a conveyer belt. Cho Oyu, on the other hand, does not face any issues of this nature.

I take advantage of the differences across these two peaks to compare the effect of language information density on the number of days it takes to reach the summit. Given the conveyer belt nature of Ama Dablam, we should expect to find no association between language information density on speed to the summit on this peak. However, if there is an association between language information density and speed of performance, the open space of strategies available while climbing Cho Oyu should allow us to observe this effect. The sample for each peak is composed of all monolingual expeditions where at least one expedition member reached the summit. The outcome variable is continuous, and in both peaks the distribution of the outcome variable was normally distributed. To test whether language information density is associated with greater speed of success, I used an Ordinary Least Squares regression model.

To test hypothesis 3, examining whether the hierarchy level of a team moderates the relationship between language information density and team performance, I use the same ZINB modeling strategy as I did for hypothesis 1 above. However, to capture the effect of team hierarchy, I delimit the analysis to small and moderately sized groups where the social interactional dynamics of leader status and expertise on intra-team communication are expected to play out. Scholars consider teams of size 10 to be the upper-bound for effective intra-team communication, with larger teams leading to fewer people talking (Thompson 2008) and to the creation of sub-groups and cliques which are then likely to engender different kinds of interactional and communication dynamics (Dunbar 1993). Commonly observed group sizes for task and decision-making groups rarely go above 15 members, with production teams averaging eight team-members (Bunderson 2003), juries ranging from six to fifteen individuals (Leib 2007), U.S. House and Senate subcommittees ranging between ten and fifteen members (Haas 2018), and central bank policy boards being composed of fewer than seven individuals (Galesic et al. 2012). The team hierarchy variable can then be substantively operationalized for small to moderately sized teams of up to ten or fifteen individuals.

Results

Descriptive statistics are presented in table 3.1, and correlations are presented in table 3.2. The average team size was 5.7 individuals (SD = 4.8), and an average of 1.2 expedition members reached the summit (SD = 2.2). The mean speed to the summit was 19.6 days (SD = 7.2) on Cho Oyu and 9.7 (SD = 5.7) days on Ama Dablam. Despite some of the control variables being highly correlated, which could raise concerns about multicollinearity, the mean variance inflation factor (VIF) did not exceed 10 in any of the models, suggesting that collinearity was not a significant issue (Belsley et al. 1980).

Table 3.1: Descriptive Statistics

Variable Type	Variable	Obs	Mean	Std. Dev.	Min	Max
DV	Count of summit members	1,626	1.15	2.24	0.00	39.00
	Proportion of summit members	1,626	0.22	0.33	0.00	1.00
	Speed to summit Cho Oyu	160	19.63	7.23	0.00	36.00
	Speed to summit Ama Dablam	155	9.66	5.72	0.00	38.00
IV	Linguistic information density	1,626	48.63	24.83	0.00	100.00
Environmental Characteristics	Region					
	Season					
	Year	1,626	1997	13	1907	2015
Risk Preferences	Terminated because too risky (1=yes)	1,626	0.57	0.50	0.00	1.00
	Std. route dummy (1=yes)	1,626	0.43	0.49	0.00	1.00
	Illegal route dummy (1=yes)	1,626	0.01	0.12	0.00	1.00
Expedition Attributes	Team size	1,626	5.70	4.81	2.00	76.00
	Mean age	1,626	36.32	7.09	20.67	70.00
	Hired non-sherpas	1,626	0.10	1.41	0.00	32.00
	Mean leader experience	1,626	3.84	4.51	1.00	45.00
	Mean member experience	1,626	2.58	2.33	1.00	22.50
	Std. dev. member experience	1,626	1.47	2.24	0.00	24.75
	Percent female	1,626	0.09	0.17	0.00	1.00
	Number of camps	1,626	1.97	1.50	0.00	8.00
	Number of climbers used oxygen	1,626	0.47	2.96	0.00	59.00
	Peak height in meters (log)	1,626	8.95	0.10	8.64	9.09
Culture attributes	Hierarchy index	1,626	-0.46	1.07	-1.96	2.13
	Mastery	1,626	3.92	0.15	3.66	4.41
	Harmony	1,626	4.18	0.35	3.46	4.62
	Embeddedness	1,626	3.45	0.21	3.10	4.18
	Individualism	1,626	62.85	19.12	12.00	91.00
	Masculinity	1,626	60.84	19.58	5.00	110.00
	Uncertainty avoidance index	1,626	73.06	20.71	23.00	112.00
Home Country Attributes	Gini coefficient	1,626	34.56	4.80	25.64	57.43
	GDP per capital	1,626	16234	6101	1245	32106
	Population size (log)	1,626	7.81	0.42	6.37	9.03
	Climatic demands index	1,626	71.26	13.52	35.00	101.00
	Education index	1,626	0.75	0.07	0.51	0.91

Table 3.2: Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1 Count of summit members																												
2 Proportion of summit members	0.63																											
3 Language information density	0.15	0.06																										
4 Terminated because too risky (1=yes)	-0.59	-0.76	-0.09																									
5 Std. route dummy (1=yes)	0.07	0.04	0.20	-0.11																								
6 Illegal route dummy (1=yes)	0.05	0.14	0.01	-0.11	-0.05																							
7 Year	0.02	0.06	0.19	-0.03	0.31	-0.13																						
8 Team size	0.52	-0.08	0.08	-0.08	0.00	-0.04	-0.15																					
9 Mean age	-0.08	-0.03	0.10	-0.02	0.20	-0.07	0.40	-0.11																				
10 Hired non-sherpas	0.32	0.02	0.15	-0.08	0.08	-0.01	0.07	0.46	-0.05																			
11 Mean leader experience	0.07	0.11	0.03	-0.11	0.07	-0.05	0.25	-0.04	0.31	0.12																		
12 Mean member experience	0.06	0.16	0.04	-0.10	0.08	-0.03	0.25	-0.14	0.35	0.06	0.81																	
13 Std. dev. member experience	0.06	0.10	0.04	-0.10	0.07	-0.06	0.25	-0.04	0.30	0.07	0.87	0.76																
14 Percent female	-0.03	-0.01	-0.03	0.02	0.02	-0.01	0.11	-0.04	0.10	-0.02	0.05	0.05	-0.01															
15 Number of camps	0.08	0.02	-0.02	-0.07	0.28	-0.05	-0.12	0.20	-0.05	-0.06	-0.10	-0.09	-0.08	-0.01														
16 Number of climbers used oxygen	0.62	0.07	0.19	-0.13	0.14	-0.01	0.06	0.69	-0.05	0.69	0.08	0.06	0.07	-0.02	0.04													
17 Peak height in meters (log)	-0.01	-0.14	0.16	0.05	0.67	-0.08	0.14	0.14	0.10	0.08	0.06	0.10	0.04	0.00	0.42	0.17												
18 Hierarchy index	0.20	0.05	-0.09	-0.06	-0.10	-0.02	-0.07	0.21	-0.08	0.18	-0.01	-0.01	0.00	0.01	0.04	0.24	-0.07											
19 Mastery	0.22	0.05	0.00	-0.07	-0.10	-0.04	-0.12	0.23	-0.13	0.23	0.00	0.02	-0.01	-0.01	-0.04	0.30	-0.11	0.68										
20 Harmony	-0.06	-0.02	-0.36	0.00	0.13	0.01	0.04	-0.04	0.06	-0.08	0.01	0.00	0.00	-0.05	0.08	-0.12	0.12	-0.61	-0.58									
21 Embeddedness	0.12	0.07	0.32	-0.08	0.15	-0.02	0.12	0.09	0.07	0.10	0.03	0.06	0.04	0.00	0.11	0.15	0.17	0.61	0.43	-0.56								
22 Individualism	-0.21	-0.06	0.20	0.09	-0.23	0.00	-0.08	-0.22	0.03	-0.16	-0.01	-0.02	0.00	0.06	-0.20	-0.22	-0.23	-0.34	-0.11	-0.31	-0.20							
23 Masculinity	0.02	-0.01	-0.58	-0.01	-0.18	-0.03	-0.18	0.08	-0.07	0.02	0.01	0.01	-0.01	-0.01	0.00	0.02	-0.14	0.45	0.50	-0.04	0.20	-0.04						
24 Uncertainty avoidance index	-0.09	-0.02	-0.55	0.03	0.07	0.01	0.04	-0.06	0.04	-0.15	0.00	-0.01	0.01	-0.01	0.17	-0.17	0.13	-0.03	-0.41	0.49	-0.06	-0.59	0.02					
25 Gini coefficient	0.11	0.04	0.33	-0.03	0.14	-0.01	0.06	0.07	0.04	0.12	0.03	0.10	0.00	0.02	0.04	0.18	0.14	0.12	0.36	-0.44	0.42	-0.10	-0.14	-0.21				
26 GDP per capita	-0.19	-0.06	-0.42	0.10	-0.36	0.00	-0.21	-0.19	-0.10	-0.18	-0.06	-0.08	-0.06	0.09	-0.17	-0.23	-0.34	-0.10	0.12	-0.13	-0.42	0.64	0.30	-0.30	-0.29			
27 Population size (log)	0.23	0.08	-0.10	-0.08	-0.12	-0.02	-0.09	0.20	-0.06	0.21	0.07	0.10	0.07	0.04	-0.05	0.26	-0.08	0.44	0.64	-0.49	0.20	0.05	0.38	-0.22	0.47	0.20		
28 Climatic demands index	0.11	0.08	0.71	-0.09	0.05	0.04	0.17	0.04	0.09	0.06	-0.01	-0.05	0.02	0.02	-0.02	0.07	0.03	0.07	-0.13	-0.29	0.17	0.15	-0.52	-0.25	-0.11	-0.16	-0.13	
29 Education index	-0.23	-0.05	-0.02	0.08	-0.32	0.02	-0.12	-0.25	-0.03	-0.24	-0.09	-0.13	-0.08	0.06	-0.13	-0.30	-0.31	0.03	0.09	-0.35	-0.11	0.67	0.13	-0.35	-0.29	0.75	-0.05	0.29

Table 3.3 reports the results of the ZINB model. Variables were introduced in three steps: only the independent variable, the independent variable and expedition-level attributes, and all control variables including country-level measures. Model 3 in table 3.3 reports the results of the ZINB model with all control variables included. As hypothesized, I find that speaking higher information density languages is associated with beneficial group outcomes (b = 0.0169, $p <$ 0.001). I find that among the expeditions that had an opportunity to summit, a one-unit change in the language information density rate is associated with an increase in the expected rate of summiting of 1.7 percent, holding all other factors constant. In turn, a one standard deviation increase in language information density is associated with a 52.1 percent increase in the count of expedition members that summit, holding all other variables constant. Looking at figure 3.3, we can see that being at the bottom of the information density distribution is associated with summiting .49 members, while being at the top of the distribution is associated with summiting 2.6 members. The results presented in table 3.3 employ standard errors clustered at the country level, but the findings hold with robust standard errors as well. These results were robust to alternative model specifications such as modeling the data using a multilevel approach with expeditions nested within countries (results available upon request).

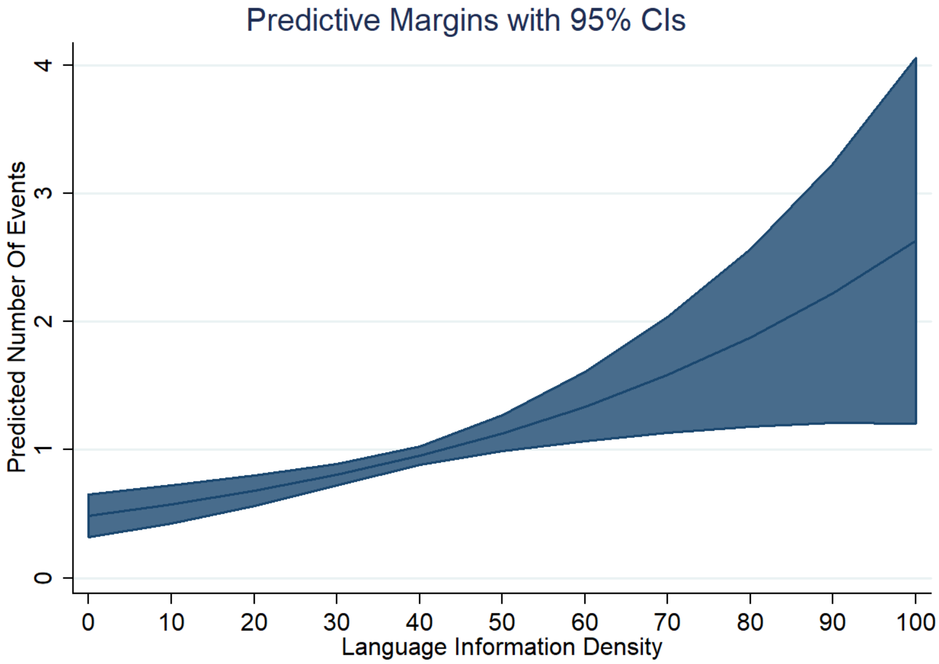
Table 3.3: Zero-Inflated Negative Binomial Model Results

VARIABLES	(1) Count of Summit Members	(2) Count of Summit Members	(3) Count of Summit Members
Language Information Density	0.00722**	0.0139***	0.0169***
	(0.00268)	(0.00335)	(0.00447)
Std. Route Dummy (1=yes)		0.231	0.284*
		(0.123)	(0.131)
Illegal Route Dummy (1=yes)		0.317**	0.267*
		(0.117)	(0.116)
Team Size		-0.949***	-0.950***
		(0.0183)	(0.0177)
Mean Age		-0.0183**	-0.0199***

		(0.00640)	(0.00598)
Hired Non-Sherpas		-0.0303*	-0.0313*
		(0.0137)	(0.0139)
Mean Leader Experience		0.0378**	0.0406***
		(0.0118)	(0.0114)
Mean Member Experience		0.0293	0.0287
		(0.0162)	(0.0169)
Std. Dev. Member Experience		-0.0652**	-0.0723**
		(0.0243)	(0.0249)
Percent Female		-0.273	-0.354
		(0.209)	(0.206)
Number of Camps		0.136***	0.124***
		(0.0380)	(0.0369)
N. of Climbers Used Oxygen		0.00468	0.00399
		(0.0123)	(0.0121)
Peak Height in Meters (log)		-4.460***	-4.692***
		(0.859)	(0.921)
Hierarchy Index		0.151	0.159
		(0.0908)	(0.0860)
Mastery		1.153*	-0.524
		(0.587)	(0.717)
Harmony		0.632	0.812*
		(0.376)	(0.342)
Embeddedness		-0.408	0.175
		(0.288)	(0.441)
Individualism		0.00420	-0.00322
		(0.00559)	(0.00623)
Masculinity		0.00179	0.00328
		(0.00353)	(0.00461)
Uncertainty Avoidance Index		0.0114**	0.00930
		(0.00384)	(0.00527)
Gini Coefficient			0.00541
			(0.0188)
GDP Per Capita			4.67e-05*
			(2.18e-05)
Population Size (log)			0.402
			(0.289)
Climatic Demands Index			0.00425
			(0.00523)
Education Index			-1.454
			(1.553)
Year		0.00390	0.00172
		(0.00332)	(0.00404)
Constant	-4.200***	24.69*	32.09*
	(0.144)	(11.25)	(14.01)
Observations	1,626	1,626	1,626
Season Fixed Effects	No	Yes	Yes
Region Fixed Effects	No	Yes	Yes

Clustered robust standard errors in parentheses (at the country level)
*** p<0.001, ** p<0.01, * p<0.05

Figure 3.3: Predicted number of expedition member summits as a function of the language information density rate.



I next tested whether the effects observed were consistent across time by subsampling expeditions from the following time periods: 1960-2015, 1970-2015, 1980-2015, and 1990-2015. I followed the same ZINB modeling strategy as above for each of these subsamples. As models 1-4 in table 3.4 indicate, all language information density coefficients are significant ($p < 0.001$) and are within the same effect size range. These results suggest that within the context of mountaineering expeditions, the effect of language structure on group behavior has been consistent across time periods. Further, because there were some expeditions with many members (max = 76), I also tested whether the effects were robust to removing larger expeditions. This is important given that the primary mechanism posited involves communication processes, which could be different across teams of different sizes, but especially very large teams. Model 5 in table 3.4 shows that language information density remains significant ($p < 0.001$) after removing all groups with more than 15 members. The results are also robust to the inclusion of expeditions to Ama Dablam (model 6 in table 3.4, *p* < 0.001) and to the removal of expeditions to Everest, which some might argue can exhibit unique crowding dynamics during certain years and seasons (model 7 in table 3.4, *p* < 0.01). Accordingly, hypothesis 1 is supported, with teams speaking informationally denser languages being associated with greater performance.

Table 3.4: Zero-Inflated Negative Binomial Model Results for Different Data Subsets

VARIABLES	(1) 1960-2015	(2) 1970-2015	(3) 1980-2015	(4) 1990-2015	(5) Team Size < 15	(6) Including Ama Dablam	(7) Not Including Everest
Language Information Density	0.0171***	0.0168***	0.0152***	0.0142***	0.0120**	0.0151***	0.0118**
	(0.00454)	(0.00439)	(0.00433)	(0.00424)	(0.00419)	(0.00335)	(0.00456)
Env. characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Risk preferences	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Expedition attributes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Home-country attributes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Culture attributes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,594	1,578	1,504	1,188	1,577	1,859	1,454
Season Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Clustered robust standard errors in parentheses (at the country level)
*** p<0.001, ** p<0.01, * p<0.05

Table 3.5 reports the results of the OLS models used to test the speed to the summit. Model 2 shows that on Cho Oyu, expeditions speaking higher information density languages were associated with summiting more quickly (*b* = -0.319, *p* < 0.001). On Cho Oyu, I find that a one-unit change in the language information density rate of an expedition is associated with a 0.319 decrease in the number of days it takes to reach the summit. This is equivalent to a one standard deviation increase in language information density being associated with reaching the summit 7 days more quickly. Looking at figure 3.4, we find that being at the bottom of the language information density distribution is associated with summiting Cho Oyu in approximately 38 days while being at the top of the distribution is associated with summiting in approximately 6.1 days. The results presented in table 3.5 employ standard errors clustered at the country level, but the findings hold with robust standard errors as well.

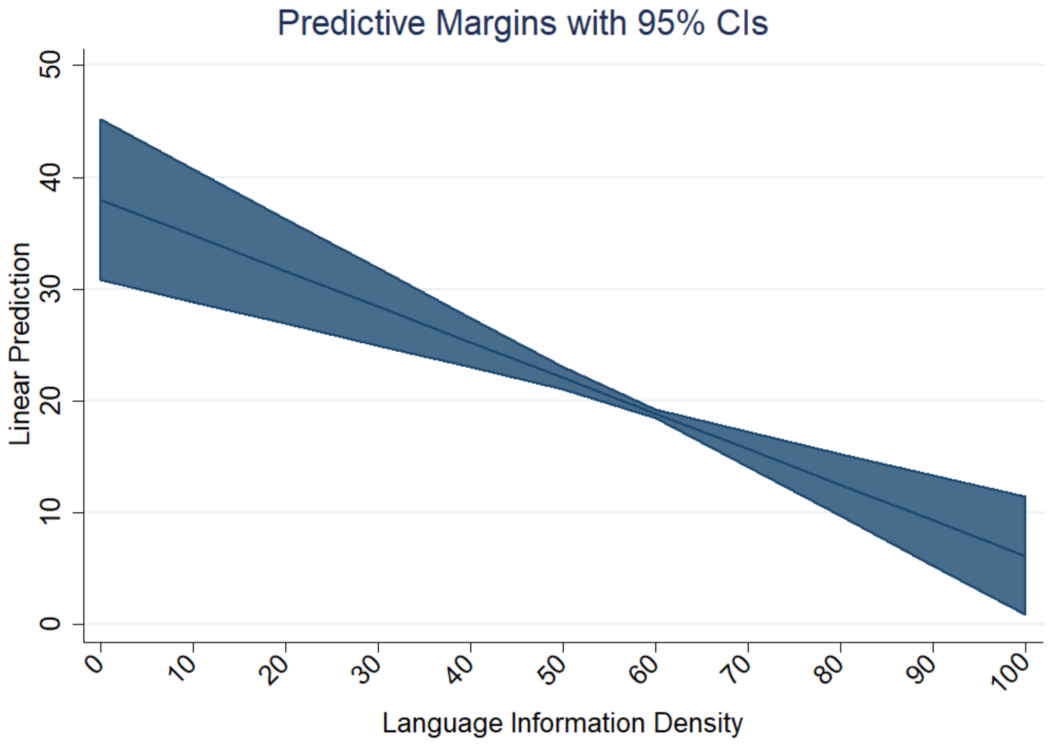
Table 3.5: OLS Regression Results

Variables	(1) Speed to Summit Cho Oyu	(2) Speed to Summit Cho Oyu	(3) Speed to Summit Ama Dablam	(4) Speed to Summit Ama Dablam
Language Information Density	-0.0909*	-0.319***	-0.0233	0.119
	(0.0363)	(0.0635)	(0.0379)	(0.0984)
Std. Route Dummy (1=yes)		-0.394		-4.992*
		(1.730)		(1.963)
Illegal Route Dummy (1=yes)		-15.79***		
		(2.674)		
Team Size		-0.0527		-0.0187
		(0.132)		(0.218)
Mean Age		-0.0955		0.0210
		(0.0872)		(0.0644)
Hired Non-Sherpas		0.470		
		(0.363)		
Mean Leader Experience		-0.286		-0.114
		(0.235)		(0.151)
Mean Member Experience		0.415		-0.0366
		(0.686)		(1.006)
Std. Dev. Member Experience		-0.278		-0.0215
		(0.547)		(0.540)
Percent Female		-6.008		1.318
		(3.312)		(2.490)
Number of Camps		0.238		1.510*
		(0.731)		(0.662)
N. of Climbers Used Oxygen		0.341		
		(0.333)		
Hierarchy Index		-2.493**		0.256
		(0.838)		(1.208)
Mastery		-31.72**		-16.28**
		(9.838)		(4.660)
Harmony		-3.769		-4.890
		(2.838)		(2.758)
Embeddedness		9.160*		-4.162*
		(4.086)		(1.778)
Individualism		-0.128		-0.131**
		(0.0740)		(0.0366)
Masculinity		-0.103***		0.0355
		(0.0266)		(0.0492)
Uncertainty Avoidance Index		-0.217*		-0.0667
		(0.0837)		(0.0522)
Gini Coefficient		0.347*		-0.637**
		(0.156)		(0.198)

GDP Per Capita		-0.000688***		-0.000279
		(0.000143)		(0.000243)
Population Size (log)		6.070*		4.747*
		(2.430)		(1.809)
Climatic Demands Index		0.111		-0.260**
		(0.0680)		(0.0679)
Education Index		70.48***		36.01*
		(16.88)		(13.52)
Year		-0.0142		-0.185
		(0.175)		(0.0907)
Constant	24.86***	98.65	10.94***	470.5*
	(2.067)	(375.7)	(2.233)	(183.9)
Observations	160	160	155	155
R-squared	0.075	0.447	0.006	0.355
Season Fixed Effects	No	Yes	No	Yes

Clustered robust standard errors in parentheses (at the country level)
*** p<0.001, ** p<0.01, * p<0.05

Figure 3.4: Predicted speed to the summit (in days) as a function of the language information density rate



Model 4 in table 3.5 shows the regression results for expeditions on Ama Dablam. As expected the coefficient for language information density was not significant. The absence of an effect was expected given the constant overcrowding on Ama Dablam that precludes teams from climbing at their own rate. As Hawley (2014) describes the 2004 Autumn season, “On Cho Oyu there were nearly twice as many teams as on Ama Dablam, but they did not have this kind of crowding problem. There was a lot more space with none of Ama Dablam’s narrow-ridge bottlenecks to confront them” (p. 341). It therefore makes sense that there is a lack of association between language information density and speed to the Ama Dablam summit, as effective communication within an expedition would have been unlikely to affect the outcome.

The significant information density coefficient on Cho Oyu, and the insignificant coefficient on Ama Dablam, provide support for H2. To be sure that this result is not an aberration in how speed to the summit was measured on each peak, I used the same modeling strategy for H1, running the ZINB model in table 3.3 on the same two subsamples (all expeditions that tried to summit Cho Oyu, and all expeditions that tried to summit Ama Dablam). Given that Ama Dablam is more of a conveyer belt to the summit rather than a desolate mountain where expeditions need to problem-solve and strategize their way to the top, we should observe the same results in the case of team members reaching the summit as in the case of speed to the summit. As expected, the results were significant on Cho Oyu (b = 0.02, p = 0.012) but were not significant on Ama Dablam (b = 0.02, p = 0.245) (results available upon request).

Finally, table 3.6 presents the results of the ZINB models testing the interaction between team hierarchy and language information density proposed in hypothesis 3. Model 1 presents the results using the subsample of expeditions with fifteen or fewer members, which was the team-size upper-bound up to which team hierarchy could be said to be substantively operationalized. The positive coefficient for the interaction between low-hierarchy teams and language information density in model 1 of table 3.6 (b = 0.007 p < 0.001) supports hypothesis 3. Figure

3.5 shows that language information density has an outsized effect on the predicted number of team-members who reach the summit for teams with low levels of hierarchy compared to teams with middle and high hierarchy levels. Table 3.6 also shows the results for subsamples including only teams with twelve or fewer members (model 2, b = 0.006, *p* < 0.001) and ten or fewer members (model 2, b = 0.005, *p* < 0.01). Running the models in table 3.6 with low-hierarchy teams as the reference category indicates that low-hierarchy teams significantly differ from both middle- and high-hierarchy teams (results available upon request). I found the same results when I ran these models by interacting language information density with quartiles of team hierarchy, with the

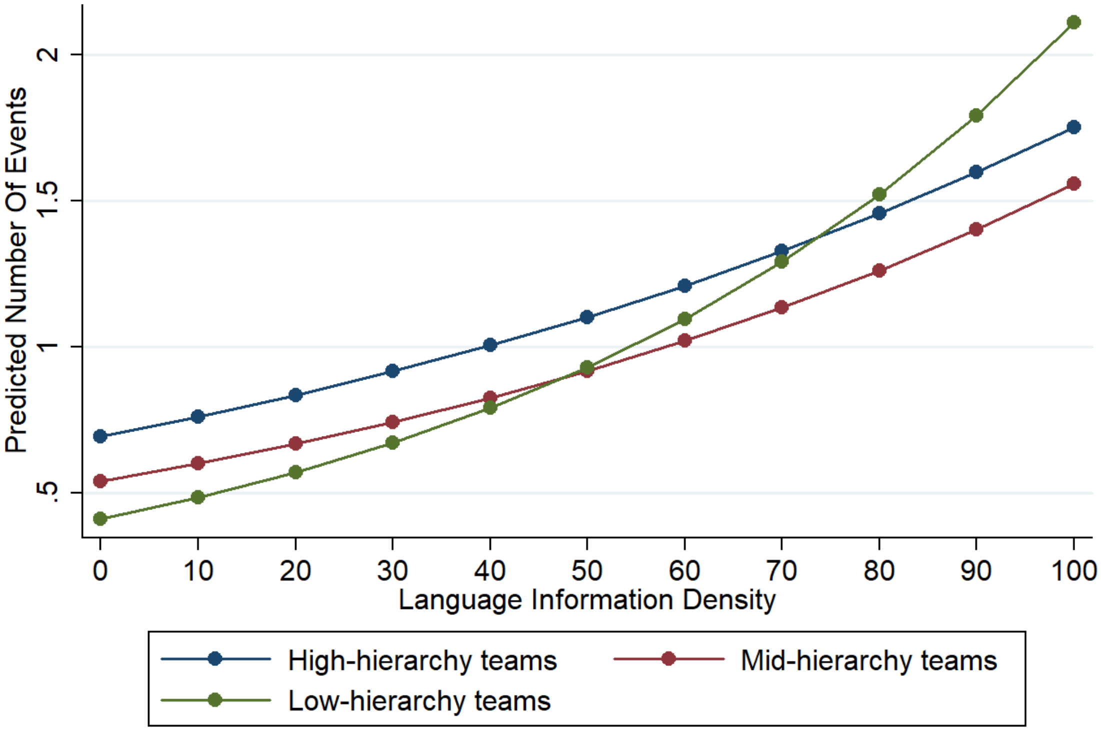
Table 3.6: Zero-Inflated Negative Binomial Model Results for Interaction Effects

VARIABLES	(1) Count of Summit Members (team size <= 15)	(2) Count of Summit Members (team size <= 12)	(3) Count of Summit Members (team size <= 10)
Language Information Density	0.00926*	0.00688	0.00564
	(0.00444)	(0.00454)	(0.00491)
Middle-hierarchy team	-0.248**	-0.240*	-0.138
	(0.0866)	(0.0984)	(0.0907)
Low-hierarchy team	-0.523***	-0.458***	-0.401***
	(0.100)	(0.0796)	(0.0955)
Mid-hierarchy teams × Language information density	0.00131 (0.00137)	0.00178 (0.00146)	0.000262 (0.00148)
Low-hierarchy teams × Language information density	0.00708*** (0.00155)	0.00612*** (0.00164)	0.00513** (0.00184)
Env. characteristics	Yes	Yes	Yes
Risk preferences	Yes	Yes	Yes
Expedition attributes	Yes	Yes	Yes
Home-country attributes	Yes	Yes	Yes
Culture attributes	Yes	Yes	Yes
Constant	33.53**	31.88***	32.54**
	(10.38)	(9.664)	(10.88)
Observations	1,577	1,542	1,472
Season Fixed Effects	Yes	Yes	Yes
Region Fixed Effects	Yes	Yes	Yes

Clustered robust standard errors in parentheses (at the country level)
*** p<0.001, ** p<0.01, * p<0.05

two least hierarchical team categories differing significantly from the most hierarchical teams in the same expected direction (*p* < 0.001 and *p* < 0.01 for each category respectively; results available upon request). In all cases, hypothesis 3 is supported. These results suggest that the mechanism shaping the association between language information density and team performance is related to group-level communication rather than the simple aggregation of individual-level cognitive effects.

Figure 3.5: Predicted number of expedition member summits as a function of the language information density rate, by level of team hierarchy**.**



Taken together, the findings above support the effect of language information density on group performance: groups speaking higher information density languages were associated with summiting a larger proportion of team members, and conditional on summiting, with doing so more quickly. Further, low-hierarchy teams where the benefits of effective communication would be most likely to be felt, were associated with increased performance effects driven by language density when compared with middle- and high-hierarchy teams.

Discussion

Language is the glue that holds social life together—the primary medium of social interaction. It is more than slightly interesting that this fact is clearly true and that this primary communication tool used by human groups has received scant attention in terms of how its morphosyntactic structure can shape the communicative constraints and possibilities of the groups that use it. The major contribution of this chapter is to bring the principle of linguistic relativity—that differences in the structure of language affect individual cognition—into sociological territory by theorizing how it is that differences in language structure can influence a group's collective cognition and performance. More than ever before, groups and organizations are living in a global, interconnected world. Research efforts have uncovered important relationships between the cultural attributes of a group and its behavior (Gelfand et al. 2017), informing cross-cultural understanding. I hope that future research will begin to generate additional insights into the role that language structure plays within group behavior. While this chapter focused on language information density, it is likely that many other language attributes influence the communicative opportunities and constraints of groups. Linguists have established variation across hundreds of structural language characteristics at the phonological, grammatical, and lexical levels (Dryer & Haspelmath 2013). Many of these are likely to play a consequential role in shaping social interaction, collective cognition, group performance, and organizational dynamics more broadly.

I combined insights from linguistics, cognitive science, and information theory through the lens of collective cognition to investigate whether the structure of the language that teams speak could affect their performance. The analysis supported this claim, showing that differences in the information density rate of a team's language is associated with differences in the group's performance outcomes. In the context of expeditions to the Himalayas, expeditions speaking higher information density languages summitted a larger number of group members, and conditional on summiting, arrived more quickly. It was also the case that low-hierarchy expeditions—those where communication would be expected to be to shape performance outcomes to the greatest degree—were associated with an outsized effect from language information density, suggesting that group communication is the likely mechanism responsible for the effects observed in this study.

Research on collective cognition has tended to emphasize processes of distributed cognition, where each group member plays a distinct role or possesses a distinct piece of the puzzle. For example, Hutchins' (1995) seminal piece on the collective cognition of ship navigation described how each member of the navigation team had expertise pertaining to their corresponding subtask, and how the group achieved the task of ship navigation by coordinating their activities in a parallel, distributed manner across individuals. Likewise, Weick and Roberts' (1993, p. 374) work on the operations of an aircraft carrier's flight deck described how "variations in contributing, representing, and subordinating produce collective mind." Similarly, work on transactive memory systems has investigated how individuals within a group encode, store, and retrieve information from distinct substantive knowledge domains (Ren & Argote 2011). In this chapter, I contribute a process of collective cognition that does not require different group members to possess different pieces of the puzzle. Even if we assume identical knowledge across group members and across language groups, it is the way that conceptual knowledge is activated by a language that is the primary mechanism proposed. This is similar to recent work on collective intelligence, which has found that communication patterns such as equality in the distribution of conversational turn-taking can spur better group performance (Woolley et al. 2010b).

A final contribution of the chapter is the use of the language information density measure. With this measure, I examined whether language structure could affect group performance in contexts requiring effective movement through conceptual space. Yet, the measure is likely to offer fruitful opportunities for theoretical development in many other domains of social activity. First, I investigated a context where group behavior was cooperative in nature. Expeditions worked together as a team to accomplish a group goal. One might wonder how information density might affect processes that are of a more conflictual nature. For example, might the process of negotiation play out differently as a function of the information density rate? One possibility is that in higher information density languages all parties to the negotiation have an expanded adjacent possible and are therefore better able to mobilize useful knowledge. Such a process could entail a more protracted and intense negotiation, generating higher rates of conflict and uncertainty, and having an uncertain effect on the outcome. Alternatively, while negotiations could be more intense in higher information density languages, a broader area of conceptual space might be activated, leading each party to reach a more optimal solution.

Second, I investigated how language information density affected the outcomes of circumscribed groups engaging in local action and communicating with each other directly. But we might wonder what effect language information density might have on larger social systems such as organizations. I theorized that higher information density would be beneficial when groups engaged in purposive collective cognition requiring effective exploration, but I wonder whether higher rates of non-purposive, unintended exploration are also possible. For example, as individuals within an organization communicate during the normal course of business, they move information throughout the organization's communication channels. As with all communication systems, an organization's communication system is also subject to noise and fidelity concerns as messages travel through. Anyone who has worked within a reasonably sized organization can recount instances where a message became distorted as it moved from person to person, not unlike the game of telephone that children play on the playground. It is likely that the expanded possibilities of meaning and interpretation possible within higher information density languages will condition more noise and less fidelity than lower information density languages. A key question is which kinds of organizational activities might benefit from the introduction of noise and loss of fidelity, and which might suffer? I suspect that activities that benefit from exploration, such as organizational innovation, might benefit from such noise, since they will be the unwitting beneficiaries of a process of unintended exploration wherein a broader area of conceptual space is covered during the normal course of activity. On the other hand, organizational activities that benefit from standardization or routinization may be worse off.

Finally, in the context of mountaineering expeditions I theorized the process of internal communication within the group. But we can also ask how the nature of mass communication might differ in languages with different information density rates. One possibility is in the context of multivocality, wherein a single message "can be interpreted coherently from multiple perspectives simultaneously… [and can] be moves in many games at once" (Padgett & Ansell 1993). My expectation would be that higher information density languages facilitate multivocal activity to a greater degree, with special consequences for political behavior both within and outside organizations. A second possibility here is in the context of social movements. Given the important role of mass communication for social movements, it is possible that higher information density languages enable some social movement tactics (e.g., the use of analogy and metaphors, multivocality) that would be less available in lower information density languages.

As with any study, this one has its limitations. The argument for the key mechanism underlying the results was the following. Individuals think through processes of associative semantic search, whereby items activated in memory elicit associated items, and each new item is added to a moving window of new possible items (Hills et al. 2015). Such a process is consistent with a biased random walk procedure, which means that the transition probabilities from one item to another shape movement through the transition probability space. As individuals converse, then, words function as cues that activate the conceptual space in memory and shape the possible directions a conversation can move in (Casasanto & Lupyan 2015). Next, I established that there are differences in the information density of languages. A consequence of higher information density within a language is a more compressed conceptual space and therefore a larger adjacent possible. In the language of biased random walks, this means that higher information density languages have more equal (i.e., higher entropy) transition probability matrices within conceptual space. The final link in the theoretical argument was that as individuals converse, a denser conceptual space as found in higher information density languages would lead to easier exploration of the space. While the logical argumentation appears sound, and while the language information density input aligned with the expected group performance output, I did not provide empirical evidence demonstrating that the information density rate of a language would lead to easier movement through conceptual space or suggesting how small or large we might expect this effect to be.

To address this concern, I created a computational simulation of biased random walks (to represent conversations) on matrices with varying transition probability distributions (to represent languages of different information density rates). Results of the computational simulation demonstrate that ease of traversal through a state space, which in our case represents a conceptual space, increases exponentially with the density of that space given a biased random walk through that space. This simulation provides mathematical validation for the expectation that higher information density languages, which have a larger conceptual adjacent possible, will facilitate movement through conceptual space during conversation. The full description and results of the simulation can be found in Appendix B2. Future experimental studies are needed to fully document the extent of this mechanism.

A great deal of group and organizational activities depend on efficient processes for collective exploration. The results of this study suggest that any activity requiring groups to engage in efficient search, creativity, problem-solving, and decision-making has the potential to be affected by the information density rate of the language being spoken. Language information density, of course, is only one of many potentially consequential language structures. Overall, this study suggests an exciting new avenue for organizational research. The fact that social life is, at its core, communicative life, implies that many fields of organizational research might be informed by integrating language structure as a central explanatory tool. This should be of interest to scholars for two reasons. First, text data, and especially social interactional text data, is becoming ever more abundant and available. Furthermore, computational techniques for the analysis of text data are growing exponentially (Evans & Aceves 2016). Without text data or computational tools, it would not have been possible to write this chapter ten years ago. This dual fact of more readily available data and computational methods for its analysis beg us to engage with language in a serious manner to inform our knowledge of organizational behavior. Second, the idea of a linguistic relativity of social interaction, collective cognition, and group performance presents itself as a potential blue ocean for organizational research. As the primary conduit of information and medium of interaction, the structure of language shapes the structure of society.

4 Moving Forward

An understanding of the role that language structure might play in shaping social interaction, collective cognition, and group performance will benefit from experimental studies designed to triangulate findings and to trace social interactional dynamics in finer detail. With these goals in mind, I have designed and begun to implement two experimental studies intended to trace how variation in language information density might influence many kinds of tasks and social interactional processes.

I expect that higher information density languages will have a number of effects on social interaction that increase the likelihood of better creativity outcomes. Individuals working in a group addressing a creativity task need to optimize their coverage of the group-level conceptual space. The structure of the language is likely to drive key social interaction dynamics that ensue within the group. If the ideal team is one that will quickly saturate the high-payoff possibilities, then this team is likely to be using a high information density language. This is because every time an individual puts forward a new idea, that new idea will be accompanied by a new set of cues, each of which will have its own associative possibilities (Abbott et al. 2015, Davelaar & Raaijmakers 2012, Hills et al. 2015). This will lead to more creative group output because the group's composite landscape will be more fully and efficiently traversed (Pirolli 2007). The effect of this process should be observable in the solution output of teams. Therefore:

> ***H1:*** *Groups that speak higher information density languages will generate <u>more</u> creative ideas compared to groups that speak lower information density languages.*
>
> ***H2:*** *Groups that speak higher information density languages will generate <u>better</u> creative ideas compared to groups that speak lower information density languages.*

The expectation of the effect of information density on creative forecasting—the skill of predicting the outcomes of new ideas (Berg 2016)—is less clear. On the one hand, such group tasks might be more likely to require negotiation among team members in order to reach a single decision judgment. This is because the normative structure of individual accountability within a group might require the possibility of assigning blame for poor choices and credit for good ones (Stewart et al. 2012). Thus, clarification may be required to have the possibility of blame and credit assignment in the future, and could lead to worse outcomes due to the increased distance that needs to be travelled to find the parts of the conceptual landscape that are necessary for agreement. On the other hand, it is possible that team members can be multi-vocal in their claims, with each member choosing to interpret what other team-members say with reference to their own interests (Padgett & Ansell 1992). Such teams will be able to efficiently search the solution space for the best outcomes while avoiding conflict. Therefore:

> ***H3a:*** *Groups that speak higher information density languages will generate worse creative forecasts because of the inefficient claims-making process compared to teams that speak lower information density languages.*
>
> ***H3b:*** *Groups that speak higher information density languages will generate better creative forecasts because of multivocality compared to teams that speak lower information density languages.*

Finally, language structure is also influential at the organizational level, where more diffuse interactions take place over the course of days and weeks. In these contexts, groups—especially multifunctional teams spread over many divisions—often interact over extended communication chains that include members spread throughout the organization. When messages travel across an organization, similar to how a message would travel during a game of telephone, there will tend to be a distorting effect to the information contained in the message due to mental and environmental noise. I argue that high information density languages will have two effects on messages. First, there will be a tendency toward greater levels of distortion, whereby the final message in a chain looks different than the initial message (Brashears & Gladstone 2016). This is because the compressed nature of the conceptual landscape will facilitate meandering through the space. Therefore:

> ***H4:*** *Groups that speak higher information density languages will generate messages at the end of a communication chain that have more diverse information compared to groups that speak lower information density languages.*

Second, the final messages will contain not only more diverse content, but content that is located across a wider distance of the conceptual space. That is, within the conceptual landscape, high information density languages will tend to randomly traverse wider areas of the landscape, leading to a much wider variation in the information content of final messages. Therefore:

> ***H5:*** *Groups that speak higher information density languages will generate messages at the end of a communication chain with a much wider set of informational content compared to groups that speak lower information density languages.*

I describe the research design used to examine these expectations below.

Creativity, Judgment, Moralizing, and Cooperation in Telugu, English, and Hindi

The first study aims to understand whether and how the information density rate of a language might shape the nature of communication processes, social interaction dynamics, and ultimate group outcomes in diverse tasks. This study is currently underway at the Center for Experimental Social Science in Pune, India, a joint Center between Oxford University and FLAME University, and is funded by a National Science Foundation Dissertation Research Improvement Grant. An advantage of carrying out this research in India is the wide diversity of languages spoken there and occupying locations at varying points on the information density distribution. A further advantage is the fact that there are many multilingual speakers who have the same cultural and educational backgrounds, including equal knowledge of at least two languages. This allows for the randomization of similar individuals to groups assigned to speak only one of these languages. Participants for this study are drawn from Telugu-English and English-Hindi bilinguals. The information density rate for Telugu is -0.85 (high information density), for English it is -1 (middle information density), and for Hindi it is -1.13 (low information density). These differences capture a large section of the information density variation across the world's most spoken languages. For each set of bilinguals, 360 individuals will be recruited and randomized 120 three-person groups assigned to speak only one of the languages. Thus, the study will be composed of 720 total individuals randomized into 240 total groups. Much of the research procedure outlined below is borrowed and adjusted from research conducted by Woolley et al. (Woolley et al. 2010b), which measures group intelligence through the accomplishment of team tasks.

When participants arrive at the laboratory, they will begin by taking a revealed preferences survey to determine their language abilities, after which they will complete a brief survey that includes intelligence and personality measures before getting together with their group for the team tasks. After this survey, teams will begin work on team tasks. This work will take place in a private laboratory room equipped with video cameras and microphones. After the group work, individuals will fill out a survey requesting information about their experience with their teams. The study will conclude with the group members playing a public goods game with each other. While recruiting and consent be carried out in English, laboratory procedures will take place in either English, Hindi, or Telugu.

To measure subjects' language abilities, I pursued a revealed preference approach. The intuition is that subjects have more accurate information about their language abilities than can be provided by any exam that is feasibly conducted in a short time period. As subjects arrive for the experiment, they are asked about preferences for the two languages they speak. The experimenter explains to the subjects that they will be required to present on a random topic of the experimenter's choosing for one or two minutes in one of the languages of their choosing (the options are Hindi and English or Telugu and English, depending on the subject pool). There is no more information about the potential topic, so that the subjects must think about their language abilities as a whole and not base their assessment on specific topical contexts. The experimenter will then tell the subjects that their performance will be evaluated from one to ten by an outside jury, and that they will be paid INR 10 for each point (INR 100 being the maximum payoff). Given that the subjects are not presented with any information about the topic, I expect them to choose their dominant language to maximize their potential payoff.

After subjects have chosen their preferred language, they are faced with a series of choices resembling a "multiple price" list (Holt & Laury 2002) and widely used in the experimental economics literature (Andersen et al. 2006; Croson 2005). Subjects are presented a list of ten "yes/no" questions of the form "Would you change your decision to the other language if you are assured an extra payoff of INR X" where X varies in INR 10 increments from 10 to 100. This method allows us to find the inflection point at which subjects would prefer to carry out the task in their less dominant language. The lower the extra amount they are willing to accept to switch to the less dominant language, the closer I can infer the subject's abilities in both languages. If a subject is willing to switch only at a high price, we infer that there is a bigger gap in the subject's abilities between the languages.

The pre-survey will include several items related to intelligence and personality traits. Participants will be given 10 minutes to complete half of the Raven's Advanced Progressive Matrices (RAPM) test, which generates a measure of general fluid reasoning capacity. Each item presents a 3 x 3 array of shapes with the lower-right corner empty. Based on patterns in the sequences of shapes, subjects must pick which of 8 other shapes properly belongs in the empty space. There is only one correct answer per question. Questions become progressively more difficult as the test goes on. Participants will next have five minutes to take the Big Five Inventory, a widely used personality test that measures the five primary personality domains: extraversion, agreeableness, conscientiousness, openness to experience, and neuroticism. They will then have fifteen minutes to take the "Reading the Mind in the Eyes" test (Baron-Cohen et al. 2001), which measures the degree to which an individual is able to attribute mental states to oneself or another person. Participants are presented with 35 photographs of the eye region of many different faces of both genders. For each photograph, they are asked to choose one of four words that best describe what the person in the photograph is thinking or feeling (e.g. frustrated, lustful, worried). The test provides a discerning measure of the degree to which individuals in a team can infer the thoughts and feelings of others. Next, participants will be asked to distribute points between themselves and an imaginary person as described in Murphy et al. (2011). This exercise measures the social value orientation of individuals, which is the magnitude of concern people have for others during interdependent decision-making. Participants will then spend seven minutes on an individual idea generation task where they will be asked to generate as many novel uses for a pillow and a brick as possible. Finally, participants will be asked to provide judgement estimates on a variety of questions such as estimating the proportion of the Indian population that responded "Very Good" in a nationally representative survey in response to questions about the influence of national institutions on the country.

After the individual measures have been collected, individuals will join their other group members in a room for the group tasks. First, groups will work together and be given 15 minutes to complete the idea generation task like that described by Pearsall and Evans (2008). Teams will be told to generate as many ideas for a blanket as possible. Teams will then be asked to choose their best idea and to turn it into a product that they can sell. They will also be told that their ideas will be rated according to two criteria: (1) the total number of ideas generated, and (2) the originality and novelty of those ideas. They will also be told that the best-performing team will be awarded a 6000 INR ($100 dollar) prize to be shared evenly among team members. Next, teams will have ten minutes to answer judgment questions such as the ones in the section on individual judgments above. They will also be told that the best-performing team will be awarded a 6000 INR ($100 dollar) prize to be shared evenly among team members.

Next, groups will be given ten minutes to discuss answers to two moral dilemmas as used in research on moral decision making (e.g., Greene et al. 2009). As an example, the first question reads as follows: “Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables. Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death. Is it appropriate for you to smother your child in order to save yourself and the other townspeople, yes or no?” After the two moral dilemmas are presented and answered, groups will then be given two minutes to discuss whether they would like to change their answer to the first dilemma. At the end of the team tasks, individual group members will be asked to answer a survey about their experience with their team. This survey will include questions related to group satisfaction (Wageman et al. 2005), motivataion (Wageman et al. 2005), social cohesiveness (Stokes 1983), and psychological safety (Edmondson 1999).

As a final task, participants will be asked to play 12 rounds of the popular public goods game that is used in economics research (Fehr & Gachter 2000; Gintis 2000). Individuals will be given an endowment of 10 points each round to use to make a joint investment with the group. Each player decides how much he/she contributes to the investment. This will be their round contribution. Each round, their contribution can be any amount from 0 points to 10 points. Their round total will be calculated as follows. In each round, the contributions of all 3 members are added up and the total is then multiplied by 2. After being multiplied by 2, the new total is evenly split among the 3 members in that round. This will be their individual share. At the end of the rounds, they will get to keep their individual share, plus the amount they did not contribute at the start of each round. There will be 12 rounds and the entire process will take a few minutes.

This study will generate approximately 150 conversation hours of transcribed text along with the associated video of the group and of each individual during the group tasks. There will be one general camera capturing the entire group conversation as well as individual laptop cameras in front of each individual capturing their individual facial mannerisms.

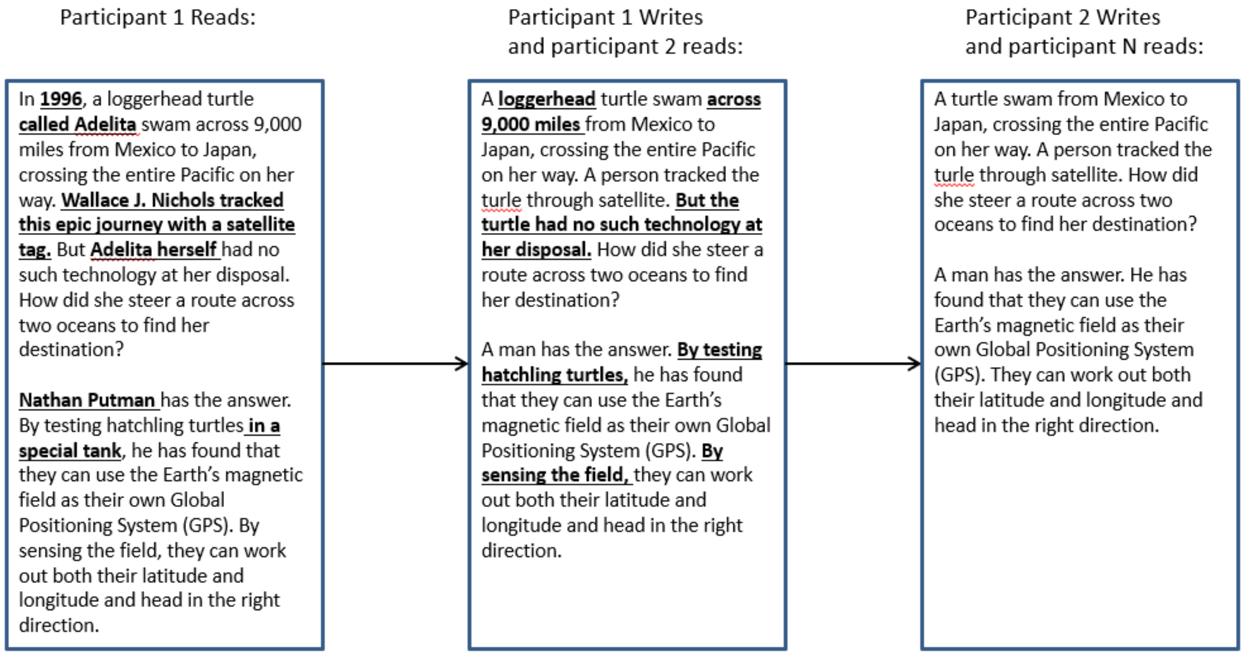
Flowing Through Conceptual Space in Hebrew, Chinese, English, and Hindi

The structure of language is also expected to be influential beyond the immediate social interactional level, where more diffuse interactions—such as across an organization—take place over the course of days and weeks. The second study seeks to determine the role that language information density has on the transmission of messages throughout a network. In these diffuse contexts, individuals often interact over extended communication chains that include members distributed across physical, geographic, and temporal space. When messages travel across such space, there is inevitably a certain degree of distortion to the information contained in the message due to mental and environmental noise that gets introduced.

I expect that higher information density languages will have two effects on the information content of these messages. First, there will be a tendency toward greater levels of distortion, whereby the final message in a chain looks more different than the initial message. This is because the density of the conceptual space will on average permit more movement within the space. Groups speaking higher information density languages will be expected to generate messages at the end of a communication chain that differ from messages at the start of the communication chain to a greater degree compared to groups that speak lower information density languages. Second, the final messages in higher information density languages will contain not only more different content, but content that has much wider variation in meaning. That is, within the dense conceptual space, groups speaking high information density languages will tend to randomly traverse wider areas of the space, leading to a much wider variation in the meaning content of final messages. Therefore, groups speaking higher information density languages will tend to generate messages at the end of a communication chain with a much wider set of informational content compared to groups speaking lower information density languages.

I follow a procedure similar to Brashears and Gladstone (Brashears and Gladstone 2016) and Hunzaker (2016), who investigated how messages accumulate errors as they diffuse through a network. This study will include 30 groups composed of five members in each of four distinct languages (Hebrew (highest), Chinese (high), English (middle), and Hindi (low)), leading to a total of 600 participants. Individuals will begin by reading a paragraph story (see figure 4.1 below) and will then be asked to take an intelligence test. After the test, individuals will then be asked to re-write the story in order to replicate the original message as closely as possible. The ending message that the first person writes will then serve as the starting message for the next person along the chain. The chain will end after five individuals have written their stories. I will trace how messages discard old information and incorporate new information along the chain of diffusion. The study will be carried out by Survey Sampling International (SSI), a leading global supplier of telephone and online surveys. I will trace how messages incorporate errors along the chain of diffusion through the use of information theoretic measures of mutual information as well as through distance measures generated from the word-embedding models described above.

Figure 4.1: Temporal flow of the telephone experiment.



Participant 1 Reads:

In **1996**, a loggerhead turtle **called Adelita** swam across 9,000 miles from Mexico to Japan, crossing the entire Pacific on her way. **Wallace J. Nichols tracked this epic journey with a satellite tag.** But **Adelita herself** had no such technology at her disposal. How did she steer a route across two oceans to find her destination?

Nathan Putman has the answer. By testing hatchling turtles **in a special tank**, he has found that they can use the Earth's magnetic field as their own Global Positioning System (GPS). By sensing the field, they can work out both their latitude and longitude and head in the right direction.

Participant 1 Writes
and participant 2 reads:

A **loggerhead** turtle swam **across 9,000 miles** from Mexico to Japan, crossing the entire Pacific on her way. A person tracked the turle through satellite. **But the turtle had no such technology at her disposal.** How did she steer a route across two oceans to find her destination?

A man has the answer. **By testing hatchling turtles,** he has found that they can use the Earth's magnetic field as their own Global Positioning System (GPS). **By sensing the field,** they can work out both their latitude and longitude and head in the right direction.

Participant 2 Writes
and participant N reads:

A turtle swam from Mexico to Japan, crossing the entire Pacific on her way. A person tracked the turle through satellite. How did she steer a route across two oceans to find her destination?

A man has the answer. He has found that they can use the Earth's magnetic field as their own Global Positioning System (GPS). They can work out both their latitude and longitude and head in the right direction.

Appendix A: Materials and methods for the estimation of language information density

Data Sources

The Bibles were gathered from two sources, bible.com and christos-c.com/bible/ (Christodouloupoulos & Steedman 2015). The bible.com documents were manually scraped with the use of a computer program while the christos-c.com documents were downloaded. The European Parliamentary Proceedings corpus was downloaded from statmt.org/europarl/ (Koehn 2005). The audio Bible duration data was manually scraped from metadata of the MP3 files found at wordproject.org and faithcomesbyhearing.com. The audio Universal Declaration of Human Rights duration data was manually scraped from metadata of the MP3 files found at udhr.audio.

Pre-processing Procedure

Four languages required the use of algorithms for word-level text segmentation. For Chinese I used the jieba segmentation package for python found at github.com/fxsjy/jieba. For Japanese I used the TinySegmenter package found at github.com/SamuraiT/tinysegmenter. For Thai I used the PyThai package found at github.com/PyThaiNLP/pythainlp. Finally, for Vietnamese I used the pyvi package found at github.com/trungtv/pyvi. Additional pre-processing for all documents included lowercasing, removing punctuation symbols, and removing numerals.

Appendix B1: Quotes from Hawley's *Seasonal Stories* regarding Ama Dablam

Year	Quotes from Hawley's Seasonal Stories (2014)
1990	"On Ama Dablam and Everest there was overcrowding in the base-camp areas, and on the smaller mountain, some teams had to wait for days while others finished climbing its very narrow southwest ridge" (p. 82).
1992	"Nearby Pumori and Ama Dablam were similarly overloaded this autumn with climbers crowding the same routes. Some Ama Dablam climbers from seven expeditions had to wait several days for their turn to try for the summit because of the lack of camping space on its narrow southwest ridge" (p. 122).
1993	"The crowding by a number of climbers on this ridge [southwest ridge of Ama Dablam] at the same time has become almost as great as that on Everest's southeast ridge, and even in winter there can be problems" (p. 126).
1998	As the season began, the Nepalese authorities expected 16 teams to come to Ama Dablam, but they continued to grant permits to everyone who asked for them, and by the time autumn ended, an all-time high number of 30 teams had been there… Furthermore, all of this autumn's expeditions had chosen to climb the same route up southwest ridge, which is quite narrow in some sections… Several people remarked that base camp looked like a carnival" (p. 246).
2004	"On Cho Oyu there were nearly twice as many teams as on Ama Dablam, but they did not have this kind of crowding problem. There was a lot more space with none of Ama Dablam's narrow-ridge bottlenecks to confront them" (p. 341).

Appendix B2: Mathematical Simulation of Movement Through Conceptual Space

With this simulation I provide an existence proof (Axelrod 1997; Miller & Page 2009) that small differences in the distribution of transition probabilities within a state space lead to significant shifts in the efficiency of movement through that space. Using a simulation of mean first passage times for Markov chain biased random walks on matrices with varying transition probabilities (Redner 2001; Sheskin 1995), I formalize the idea that more equal (i.e., higher density, higher entropy) transition probabilities within a state space lead to easier traversal through the space. A simulation such as this is a way of answering the question: starting from a random problem state *i*, how many steps will it take, on average, to reach a random solution state *j*, given the underlying transition probability distribution of the space? Given an *n* by *n* matrix where each cell contains the transition probability from row *i* to column *j*, the mean first passage time is the mean number of steps required to go from state *i* to state *j* for the first time (Redner 2001; Sheskin 1995). Translated into the language of conceptual spaces above, this simulation models how groups are likely to move differently through conceptual space depending on how the information density rate of a language has shaped the transition probabilities between concepts.

Mathematically, the mean first passage time from state *i* to state *j* is denoted by m_{ij}. This is the expected number of transitions before reaching state *j*, given we are starting at state *i*. We can move from state *i* to state *j* in two ways. First, we could move directly from state *i* to state *j* in one step with probability p_{ij}. Second, we could move from state *i* to state *k* in one step with probability p_{ik}, so long as $k \neq j$. Now we are in state *k*. It will then take m_{kj} steps to go from *k* to *j*. It will therefore take an average of $1 + m_{kj}$ steps to go from *i* to *j*. Therefore,

$$m_{ij} = p_{ij} \cdot 1 + \sum_{\mathrm{k} \neq j} p_{ik}\left(1 + m_{kj}\right) = 1 + \sum_{k \neq j} p_{ik} m_{kj}$$

I solve this equation by following the algorithmic procedure in Sheskin (1995).

The first step in this simulation was to create transition probability matrices with different levels of density for the transition probabilities from any state *i* to any state *j*. Given that I am modeling the transition probabilities within a conceptual space, this is a fully connected matrix (or graph), given that every concept has a non-zero probability of leading to any other concept, including itself[28]. The transition probabilities for any row *i* must add to one. Operationalizing density in information theoretic terms, the maximum density transition probability matrix is one where entropy is maximized, meaning that every transition probability is equal to 1/*n*, where *n* is the number of states. For example, in a 100x100 matrix, every transition probability would equal .01. To create the transition matrices, I drew numbers randomly in the range (0, 1) from a power distribution with positive exponent *a*-1 and normalized each row *i* to add up to one. By adjusting the exponent parameter *a*, I adjusted the density of the transition probability distribution.

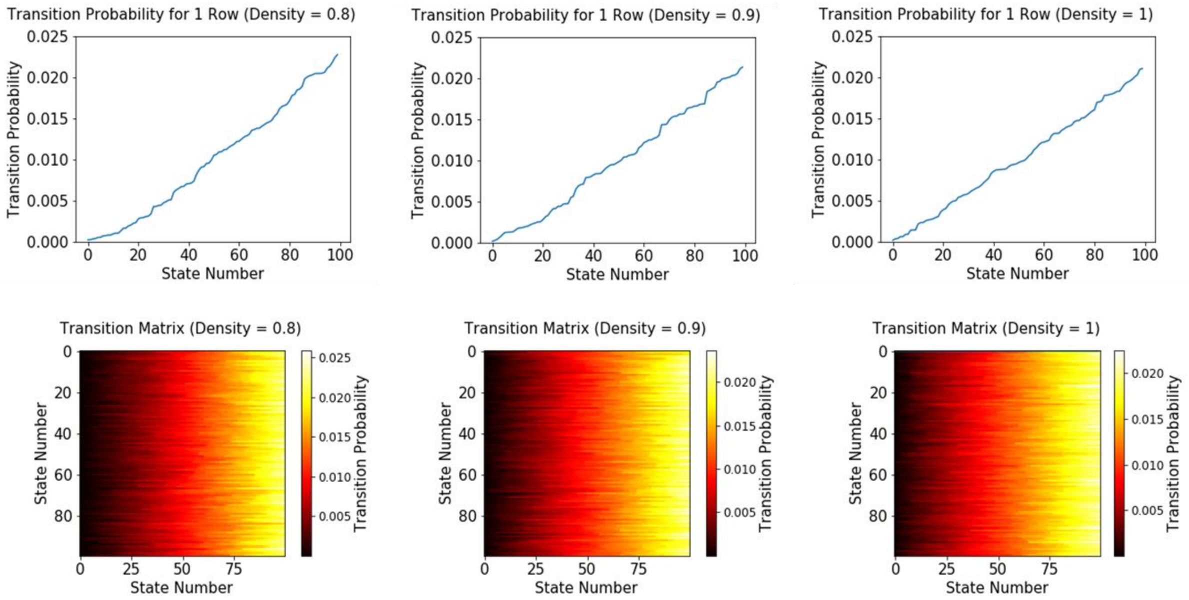
I ran simulations with transition probability matrices of 100 rows and columns[29], for values of *a* ranging from .7 (least dense) to 1 (most dense), in .02 increments. The top row in figure 1 shows the transition probability distribution for one row at each of three density levels (.8, .9, and 1). While it is not pictured, for reference we can imagine that a maximum density (or

[28] That this is true can be seen in Chomsky's classic sentence "colorless green ideas sleep furiously" (Chomsky 1957, p. 15). While semantically nonsensical, the sentence is nonetheless grammatically correct and was in fact used to communicate a message.

[29] The number of operations needed to compute the mean first passage time for each matrix is equal to $n^3/2$. This means that a 100 x 100 matrix requires 500,000 operations. While a larger matrix may have been preferable to better model a conceptual space, increasing the matrix size to only 1,000 x 1,000 would increase the number of computing operations to 500,000,000 , making it computationally infeasible. Regardless, given the underlying mathematical logic, the nature of the results would remain unchanged.

maximum entropy) distribution would be a flat line with the transition probability for each state equaling .01. A sample of matrices at differing levels of density can be found in the bottom row of figure 1. For each density level, I created 100 matrices and calculated their mean first passage time.

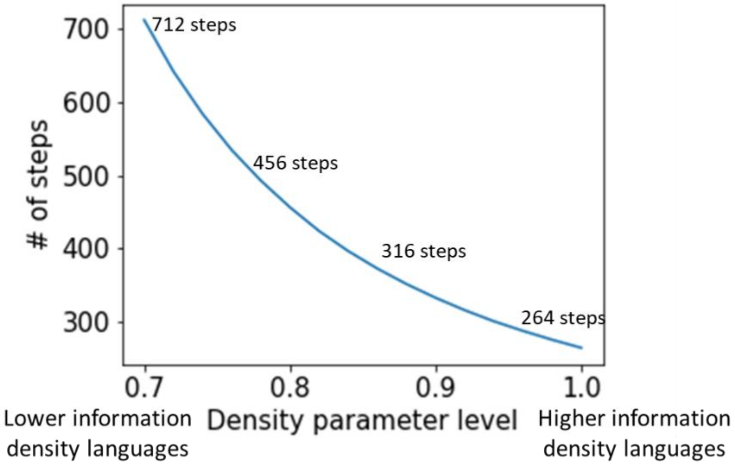
Figure 1: Shape of the transition probability distribution for one row at different degrees of density (top), and sample transition probability matrices at different degrees of density (bottom).



Results

The results of the simulated biased random walks on transition probability matrices with different densities are found in figure 2. As we can see, the average number of steps required to reach a random solution state *j* from a random problem state *i* grows exponentially as the density of the space decreases. In a 100x100 state space where the density parameter is set at 1, it takes an average of 264 steps to reach *j* from *i*. This number increases by 52 steps from a density of 1 to a density of .9, by 140 steps from a density of .9 to a density of .8, and by 256 steps from a density of .8 to a density of .7. These results demonstrate that small differences in the distribution of a transition probability matrix have significant effects on the biased random walk movement within that space, lending support to the expectation that the same process occurs as members within a group move through conceptual space during conversation. Experimental evidence will be required to fully document this expectation.

Figure 2: The average mean first passage time (in # of steps) at each density level.



References

Abbott JT, Austerweil JL, Griffiths TL. 2015. Random walks on semantic networks can resemble optimal foraging. *Psychol. Rev.* 122(3):558–69

Adamic LA, Lento TM, Adar E, Ng PC. 2014. Information Evolution in Social Networks. *ArXiv14026792 Phys.*

Altmann GTM, Kamide Y. 1999. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*. 73(3):247–64

Ancona DG, Caldwell DF. 1992. Demography and Design: Predictors of New Product Team Performance. *Organ. Sci.* 3(3):321–41

Andersen S, Harrison GW, Lau MI, Rutström EE. 2006. Elicitation using multiple price list formats. *Exp. Econ.* 9(4):383–405

Anicich EM, Swaab RI, Galinsky AD. 2015. Hierarchical cultural values predict success and mortality in high-stakes teams. *Proc. Natl. Acad. Sci.* 112(5):1338–43

Aral S, Alstyne MV. 2011. The Diversity-Bandwidth Trade-off. *Am. J. Sociol.* 117(1):90–171

Atteveldt W van, Kleinnijenhuis J, Ruigrok N. 2008. Parsing, Semantic Networks, and Political Authority Using Syntactic Analysis to Extract Semantic Relations from Dutch Newspaper Articles. *Polit. Anal.* 16(4):428–46

Axelrod R. 1997. Advancing the Art of Simulation in the Social Sciences. In *Simulating Social Phenomena*, pp. 21–40. Springer, Berlin, Heidelberg

Bail CA. 2012. The Fringe Effect Civil Society Organizations and the Evolution of Media Discourse about Islam since the September 11th Attacks. *Am. Sociol. Rev.* 77(6):855–79

Bail CA. 2014. The cultural environment: measuring culture with big data. *Theory Soc.* 43(3–4):465–82

Bail CA. 2016. Cultural Carrying Capacity: Organ Donation Advocacy, Discursive Framing, and Social Media Engagement. *Soc. Sci. Med.*

Baker SR, Bloom N, Davis SJ. 2013. *Measuring Economic Policy Uncertainty*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2198490

Bakshy E, Messing S, Adamic L. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*. aaa1160

Bales RF. 1950. A Set of Categories for the Analysis of Small Group Interaction. *Am. Sociol. Rev.* 15(2):257–63

Bayley R, Cameron R, Lucas C. 2013. *The Oxford Handbook of Sociolinguistics*. OUP USA

Bell ST. 2007. Deep-level composition variables as predictors of team performance: a meta-analysis. *J. Appl. Psychol.* 92(3):595–615

Bellos D. 2011. *Is That a Fish in Your Ear?: Translation and the Meaning of Everything*. Penguin Books Limited

Belsley DA, Kuh E, Welsch RE. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley

Berelson B, Lazarsfeld PF. 1948. *The Analysis of Communication Content*. Universitetets studentkontor

Berg JM. 2016. Balancing on the Creative Highwire: Forecasting the Success of Novel Ideas in Organizations. *Adm. Sci. Q.* 61(3):433–68

Berger J, Cohen BP, Zelditch M. 1972. Status Characteristics and Social Interaction. *Am. Sociol. Rev.* 37(3):241–55

Berger J, Rosenholtz SJ, Zelditch M. 1980. Status Organizing Processes. *Annu. Rev. Sociol.* 6:479–508

Blackburn J, Kwak H. 2014. STFU NOOB!: Predicting Crowdsourced Decisions on Toxic Behavior in Online Games. *Proc. 23rd Int. Conf. World Wide Web*, pp. 877–888. New York, NY, USA: ACM

Blei D, Lafferty J. 2006. Dynamic topic models. *Proc. 23rd Int. Conf. Mach. Learn.*

Blei D, Ng A, Jordan M. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022

Blei DM. 2012. Probabilistic topic models. *Commun. ACM*. 55(4):77–84

Bock K, Carreiras M, Meseguer E. 2012. Number meaning and number grammar in English and Spanish. *J. Mem. Lang.* 66(1):17–37

Bollen J, Mao H, Zeng X. 2011. Twitter mood predicts the stock market. *J. Comput. Sci.* 2(1):1–8

Bonilla T, Grimmer J. 2013. Elevated threat levels and decreased expectations: How democracy handles terrorist threats. *Poetics*. 41(6):650–69

Bonner BL, Baumann MR, Dalal RS. 2002. The effects of member expertise on group decision-making and performance. *Organ. Behav. Hum. Decis. Process.* 88(2):719–36

Boukreev A, DeWalt GW. 1999. *The Climb: Tragic Ambitions on Everest*. Macmillan

Bourdieu P. 2013. *Distinction: A Social Critique of the Judgement of Taste*. Routledge

Brashears ME, Gladstone E. 2016. Error correction mechanisms in social networks can reduce accuracy and encourage innovation. *Soc. Netw.* 44:22–35

Brown RW, Lenneberg EH. 1954. A study in language and cognition. *J. Abnorm. Soc. Psychol.* 49(3):454–62

Bunderson JS. 2003. Recognizing and Utilizing Expertise in Work Groups: A Status Characteristics Perspective. *Adm. Sci. Q.* 48(4):557–91

Burt RS. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press

Burt RS. 2004. Structural holes and good ideas. *Am. J. Sociol.* 110(2):349–99

Burt RS. 2005. *Brokerage and Closure: An Introduction to Social Capital*. Oxford: Oxford University Press

Cadilhac A, Asher N, Benamara F, Lascarides A. 2013. Grounding Strategic Conversation: Using Negotiation Dialogues to Predict Trades in a Win-Lose Game. *Proc. 2013 Conf. Empir. Methods Nat. Lang. Process.*, pp. 357–368. Seattle, Washington, USA: Association for Computational Linguistics

Campbell L. 2013. *Historical Linguistics: An Introduction*. Edinburgh University Press

Carley K. 1993. Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociol. Methodol.* 23(75–126):

Carley K. 1994. Extracting culture through textual analysis. *Poetics*. 22(4):291–312

Carley KM, Kaufer DS. 1993. Semantic connectivity: An approach for analyzing symbols in semantic networks. *Commun. Theory*. 3(3):183–213

Casasanto D, Boroditsky L. 2008. Time in the mind: using space to think about time. *Cognition*. 106(2):579–93

Casasanto D, Lupyan G. 2015. All Concepts Are Ad Hoc Concepts. In *The Conceptual Mind: New Directions in the Study of Concepts*, eds. E Margolis, S Laurence. MIT Press

Chang J, Blei D. 2010. Hierarchical relational models for document networks. *Ann. Appl. Stat.*

Charnov EL. 1976. Optimal foraging, the marginal value theorem. *Theor. Popul. Biol.* 9(2):129–36

Chen MK. 2013. The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *Am. Econ. Rev.* 103(2):690–731

Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J. 2014. Can Cascades Be Predicted? *Proc. 23rd Int. Conf. World Wide Web*, pp. 925–936. New York, NY, USA: ACM

Cheng J, Danescu-Niculescu-Mizil C, Leskovec J. 2015. Antisocial Behavior in Online Discussion Communities. *ArXiv150400680 Cs Stat*

Chomsky N. 1957. *Syntactic Structures*. Walter de Gruyter

Christodouloupoulos C, Steedman M. 2015. A massively parallel corpus: the Bible in 100 languages. *Lang. Resour. Eval.* 49(2):375–95

Clark A, Fox C, Lappin S. 2010. *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley

Coppola CADWG, Petrov S. 2015. *Improved Transition-Based Parsing and Tagging with Neural Networks*. Work. Pap.

Corman SR, Kuhn T, Mcphee RD, Dooley KJ. 2002. Studying Complex Discursive Systems: Centering Resonance Analysis of Communication. *Hum. Commun. Res.* 28(2):157–206

Coulmas F. 1998. *The Handbook of Sociolinguistics*. Wiley

Croson R. 2005. The Method of Experimental Economics. *Int. Negot.* 10(1):131–48

Csaszar FA, Levinthal DA. 2016. Mental representation and the discovery of new strategies. *Strateg. Manag. J.* 37(10):2031–49

Danescu-Niculescu-Mizil C, Lee L, Pang B, Kleinberg J. 2012. Echoes of Power: Language Effects and Power Differences in Social Interaction. *Proc. 21st Int. Conf. World Wide Web*, pp. 699–708. New York, NY, USA: ACM

Danescu-Niculescu-Mizil C, West R, Jurafsky D, Leskovec J, Potts C. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. *Proc. 22Nd Int. Conf. World Wide Web*, pp. 307–318. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee

Dascal M. 2006. Adam Smith's Theory of Language. In *The Cambridge Companion to Adam Smith*, pp. 79–111. Cambridge, UK: Cambridge University Press

Davelaar EJ, Raaijmakers JGW. 2012. Human Memory Search. In *Cognitive Search: Evolution, Algorithms, and the Brain*, eds. PM Todd, TT Hills, TW Robbins. Cambridge, MA: MIT Press

Davidoff J, Davies I, Roberson D. 1999. Colour categories in a stone-age tribe. *Nature*. 398(6724):203–4

de Saussure F. 1986. *Course in General Linguistics*. Open Court

DiMaggio P, Nag M, Blei D. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*. 41(6):570–606

Dolscheid S, Shayan S, Majid A, Casasanto D. 2013. The thickness of musical pitch: psychophysical evidence for linguistic relativity. *Psychol. Sci.* 24(5):613–21

Dryer MS, Haspelmath M, eds. 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology

Dunbar RIM. 1993. Coevolution of neocortical size, group size and language in humans. *Behav. Brain Sci.* 16(4):681–94

Enfield NJ. 2015. Linguistic Relativity from Reference to Agency. *Annu. Rev. Anthropol.* 44(1):207–24

Enfield NJ, Majid A, van Staden M. 2006. Cross-linguistic categorisation of the body: Introduction. *Lang. Sci.* 28(2):137–47

Evans JA, Aceves P. 2016. Machine Translation: Mining Text for Social Theory. *Annu. Rev. Sociol.* 42:

Fehr E, Gachter S. 2000. Cooperation and Punishment in Public Goods Experiments. *Am. Econ. Rev.* 90(4):980–94

Feist MI. 2008. Space Between Languages. *Cogn. Sci.* 32(7):1177–99

Fishman JA, ed. 1968. *Readings in the Sociology of Language*. Walter de Gruyter

Fligstein N, Brundage JS, Schultz M. 2014. Why the Federal Reserve Failed to See the Financial Crisis of 2008: The Role of “Macroeconomics” as a Sense making and Cultural Frame

Foster JG, Rzhetsky A, Evans JA. 2015. Tradition and Innovation in Scientists’ Research Strategies. *Am. Sociol. Rev.* 0003122415601618

Foulds JR, Smyth P. 2013. Modeling Scientific Impact with Topical Influence Regression. *EMNLP*, pp. 113–123. http://www.aclweb.org/anthology/D13-1012

Franzosi R. 2004. *From Words to Numbers: Narrative, Data, and Social Science*. Cambridge University Press

Freese J. 2007. Replication Standards for Quantitative Social Science Why Not Sociology? *Sociol. Methods Res.* 36(2):153–72

Fuhrman O, McCormick K, Chen E, Jiang H, Shu D, et al. 2011. How Linguistic and Cultural Forces Shape Conceptions of Time: English and Mandarin Time in 3D. *Cogn. Sci.* 35(7):1305–28

Galesic M, Olsson H, Rieskamp J. 2012. Social Sampling Explains Apparent Biases in Judgments of Social Environments. *Psychol. Sci.* 23(12):1515–23

Garfinkel H. 1967. *Studies in Ethnomethodology*. Prentice-Hall

Gavetti G, Levinthal D. 2000. Looking Forward and Looking Backward: Cognitive and Experiential Search. *Adm. Sci. Q.* 45(1):113–37

Gelfand MJ, Aycan Z, Erez M, Leung K. 2017. Cross-cultural industrial organizational psychology and organizational behavior: A hundred-year journey. *J. Appl. Psychol.* 102(3):514–29

Gentzkow M, Shapiro JM. 2010. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*. 78(1):35–71

Gibson DR. 2012. *Talk at the Brink: Deliberation and Decision during the Cuban Missile Crisis*. Princeton University Press

Gintis H. 2000. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Behavior*. Princeton University Press

Glaser BG. 1965. The Constant Comparative Method of Qualitative Analysis. *Soc. Probl.* 12(4):436–45

Glaser BG, Strauss AL. 1967. *The Discovery of Grounded Theory; Strategies for Qualitative Research*. Chicago,: Aldine Pub. Co.

Goffman E. 2005. *Interaction Ritual: Essays in Face to Face Behavior*. AldineTransaction

Goldberg A, Srivastava SB, Manian VG, Monroe W, Potts C. 2015. Fitting In or Standing Out? The Tradeoffs of Structural and Cultural Embeddedness. *Unpubl. Manuscr.*

Golder SA, Macy MW. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*. 333(6051):1878–81

Golder SA, Macy MW. 2014. Digital Footprints: Opportunities and Challenges for Online Social Research. *Annu. Rev. Sociol.* 40(1):129–52

Goncalo JA, Chatman JA, Duguid MM, Kennedy JA. 2015. Creativity from Constraint? How the Political Correctness Norm Influences Creativity in Mixed-sex Work Groups. *Adm. Sci. Q.* 60(1):1–30

Granovetter M. 1985. Economic Action and Social Structure: The Problem of Embeddedness. *Am. J. Sociol.* 91(3):481–510

Griffiths TL, Steyvers M, Blei DM, Tenenbaum JB. 2005. Integrating Topics and Syntax. In *Advances in Neural Information Processing Systems 17*, eds. LK Saul, Y Weiss, L Bottou, pp. 537–544. MIT Press

Grimmer J. 2010. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Polit. Anal.* 18(1):1–35

Grimmer J. 2013. Appropriators not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation. *Am. J. Polit. Sci.* 57(3):624–42

Grimmer J, King G. 2011. General purpose computer-assisted clustering and conceptualization. *Proc. Natl. Acad. Sci.* 108(7):2643–50

Grimmer J, Stewart BM. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Polit. Anal.* mps028

Grogger J. 2011. Speech Patterns and Racial Wage Inequality. *J. Hum. Resour.* 46(1):1–25

Gross DP. 2014. *Creativity Under Fire: The Effects of Competition on Innovation and the Creative Process*. Work. Pap., Berkeley, CA

Gulati R, Sytch M. 2007. Dependence Asymmetry and Joint Dependence in Interorganizational Relationships: Effects of Embeddedness on a Manufacturer's Performance in Procurement Relationships. *Adm. Sci. Q.* 52(1):32–69

Gumperz JJ, Levinson SC. 1996. *Rethinking Linguistic Relativity*. Cambridge University Press

Haas KL. 2018. *List of standing committees and select committees and their subcommittees of the House of Representatives of the United States together with joint committees of the Congress with an alphabetical list of the members and their committee assignments*. http://clerk.house.gov/ committee_info/scsoal.pdf

Harbison JI, Dougherty MR, Davelaar EJ, Fayyad B. 2009. On the lawfulness of the decision to terminate memory search. *Cognition*. 111(3):416–21

Hardt H. 2001. *Social Theories of the Press: Constituents of Communication Research, 1840s to 1920s*. Rowman & Littlefield Publishers

Hasan KS, Ng V. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. EMNLP*, pp. 751–762. http://www.aclweb.org/old_anthology/D/D14/D14-1083.pdf

Hawley E. 2014. *Seasonal Stories for the Nepalese Himalaya: 1985-2014*. http://www.himalayandatabase.com/downloads/EAH%20Seasonal%20Stories.pdf

Hawley E, Salisbury R. 2004. *The Himalayan Database: The Expedition Archives of Elizabeth Hawley*. Golden, Colo.: Amer Alpine Club

Hearst MA. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Comput Linguist*. 23(1):33–64

Henrich J, Heine SJ, Norenzayan A. 2010a. The weirdest people in the world? *Behav. Brain Sci.* 33(2/3):61–135

Henrich J, Heine SJ, Norenzayan A. 2010b. Most people are not WEIRD. *Nature*. 466(7302):29–29

Hills TT, Todd PM, Jones MN. 2015. Foraging in Semantic Fields: How We Search Through Memory. *Top. Cogn. Sci.* 7(3):513–34

Hinsz VB, Tindale RS, Vollrath DA. 1997. The emerging conceptualization of groups as information processors. *Psychol. Bull.* 121(1):43–64

Hinton GE, Osindero S, Teh Y-W. 2006. A fast learning algorithm for deep belief nets. *Neural Comput.* 18(7):1527–54

Hirschberg J, Manning CD. 2015. Advances in natural language processing. *Science*. 349(6245):261–66

Hofmann DA, Jones LM. 2005. Leadership, collective personality, and performance. *J. Appl. Psychol.* 90(3):509–22

Hofstadter D, Sander E. 2013. *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. Basic Books

Hofstede G. 1984. *Culture's Consequences: International Differences in Work-Related Values*. Sage

Hofstede G. 2001. *Culture's Consequences: International Differences in Work-Related Values*. Sage. 2nd ed.

Holmes J. 2008. *An Introduction to Sociolinguistics*. Pearson Longman

Holt CA, Laury SK. 2002. Risk Aversion and Incentive Effects. *Am. Econ. Rev.* 92(5):1644–55

Hong L, Page SE. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci.* 101(46):16385–89

Howard MW, Addis KM, Jing B, Kahana MJ. 2007. Semantic structure and episodic memory. In *LSA: A Road towards Meaning*, eds. KT Landauer, D McNamara, W Dennis, W Kintsch

Huffman DA. 1952. A Method for the Construction of Minimum-Redundancy Codes. *Proc. IRE*. 40(9):1098–1101

Huizinga J. 1971. *Homo Ludens: A Study of the Play-Element in Culture*. Beacon Press

Hunzaker MBF. 2016. Cultural Sentiments and Schema-Consistency Bias in Information Transmission. *Am. Sociol. Rev.* 81(6):1223–50

Hutchins E. 1995. *Cognition in the Wild*. MIT Press

Ioannidis J, Doucouliagos C. 2013. What's to Know About the Credibility of Empirical Economics? *J. Econ. Surv.* 27(5):997–1004

Ioannidis JP. 2005. Why most published research findings are false. *PLoS Med*. 2(8):e124

Iyyer M, Enns P, Boyd-Graber J, Resnik P. 2014. Political ideology detection using recursive neural networks. *Assoc. Comput. Linguist.* http://www.cs.colorado.edu/~jbg/docs/2014_acl_rnn_ideology.pdf

Jehn KA, Bezrukova K. 2004. A field study of group diversity, workgroup context, and performance. *J. Organ. Behav.* 25(6):703–29

Jehn KA, Northcraft GB, Neale MA. 1999. Why Differences Make a Difference: A Field Study of Diversity, Conflict, and Performance in Workgroups. *Adm. Sci. Q.* 44(4):741–63

Jelveh Z, Kogut B, Naidu S. 2014. Detecting latent ideology in expert text: Evidence from academic papers in economics. *Proc. EMNLP*. http://anthology.aclweb.org/D/D14/D14-1191.pdf

Jockers ML, Mimno D. 2013. Significant themes in 19th-century literature. *Poetics*. 41(6):750–69

Johnson S. 2010. *Where Good Ideas Come From*. Penguin

Jordan MI, Mitchell TM. 2015. Machine learning: Trends, perspectives, and prospects. *Science*. 349(6245):255–60

Joshi M, Das D, Gimpel K, Smith NA. 2010. Movie reviews and revenues: An experiment in text regression. *Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist.*, pp. 293–96. Association for Computational Linguistics

Jurafsky D, Martin JH. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, N.J.: Prentice Hall

Kauffman SA. 2000. *Investigations*. Oxford University Press

Kay P, Berlin B, Maffi L, Merrifield W. 1997. Color Naming Across Languages. In *Color Categories in Language and Thought*, eds. CL Hardin, L Maffi. Cambridge University Press

Kemp C, Tenenbaum JB. 2008. The discovery of structural form. *Proc. Natl. Acad. Sci. U. S. A.* 105(31):10687–92

Kilduff M, Angelmar R, Mehra A. 2000. Top Management-Team Diversity and Firm Performance: Examining the Role of Cognitions. *Organ. Sci.* 11(1):21–34

Kitayama S, Uskul AK. 2011. Culture, Mind, and the Brain: Current Evidence and Future Directions. *Annu. Rev. Psychol.* 62(1):419–49

Klingenstein S, Hitchcock T, DeDeo S. 2014. The civilizing process in London’s Old Bailey. *Proc. Natl. Acad. Sci.* 111(26):9419–24

Koehn P. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*. MT Summit. http://www.statmt.org/europarl/

Krakauer J. 1998. *Into Thin Air*. Knopf Doubleday Publishing Group

Kramer ADI, Guillory JE, Hancock JT. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci.* 111(24):8788–90

Krippner GR, Alvarez AS. 2007. Embeddedness and the Intellectual Projects of Economic Sociology. *Annu. Rev. Sociol.* 33(1):219–40

Kuhn T, Perc M, Helbing D. 2014. Inheritance Patterns in Citation Networks Reveal Scientific Memes. *Phys. Rev. X*. 4(4):041036

Kulkarni V, Al-Rfou R, Perozzi B, Skiena S. 2015. Statistically Significant Detection of Linguistic Change. *Proc. 24th Int. Conf. World Wide Web*, pp. 625–635. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee

Lai VT, Boroditsky L. 2013. The immediate and chronic influence of spatio-temporal metaphors on the mental representations of time in english, mandarin, and mandarin-english speakers. *Front. Psychol.* 4:142

Lakoff G. 2008. *Women, Fire, and Dangerous Things*. University of Chicago Press

Lambert D. 1992. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics*. 34(1):1–14

Laver M, Benoit K, Garry J. 2003. Extracting policy positions from political texts using words as data. *Am. Polit. Sci. Rev.* 97(02):311–31

Lavie D, Stettner U, Tushman ML. 2010. Exploration and Exploitation Within and Across Organizations. *Acad. Manag. Ann.* 4(1):109–55

Lee M, Martin JL. 2015. Coding, counting and cultural cartography. *Am. J. Cult. Sociol.* 3(1):1–33

Leib EJ. 2007. A Comparison of Criminal Jury Decision Rules in Democratic Countries Commentaries. *Ohio State J. Crim. Law*. 5:629–44

LePine JA. 2003. Team adaptation and postchange performance: effects of team composition in terms of members' cognitive ability and personality. *J. Appl. Psychol.* 88(1):27–39

Leskovec J, Backstrom L, Kleinberg J. 2009. Meme-tracking and the Dynamics of the News Cycle. *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 497–506. New York, NY, USA: ACM

Levinson SC. 1996. Language and Space. *Annu. Rev. Anthropol.* 25:353–82

Levinson SC. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press

Levinson SC. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press

Livne A, Simmons M, Adar E, Adamic L. 2011. The Party Is Over Here: Structure and Content in the 2010 Election. *Fifth Int. AAAI Conf. Weblogs Soc. Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2852

Long NE. 1958. The Local Community as an Ecology of Games. *Am. J. Sociol.* 64(3):251–61

Lorenz J, Rauhut H, Schweitzer F, Helbing D. 2011. How social influence can undermine the wisdom of crowd effect. *Proc. Natl. Acad. Sci.* 108(22):9020–9025

Loreto V, Servedio VDP, Strogatz SH, Tria F. 2016. Dynamics on Expanding Spaces: Modeling the Emergence of Novelties. In *Creativity and Universality in Language*, pp. 59–83. Springer, Cham

Loughran T, McDonald B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance*. 66(1):35–65

Lucy JA. 1992a. *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis*. Cambridge University Press

Lucy JA. 1992b. *Grammatical Categories and Cognition: A Case Study of the Linguistic Relativity Hypothesis*. Cambridge University Press

Lucy JA. 1997. Linguistic Relativity. *Annu. Rev. Anthropol.* 26(1):291–312

Lupyan G. 2012. What Do Words Do? Toward a Theory of Language-Augmented Thought. In *The Psychology of Learning and Motivation*, ed. BH Ross. Academic Press

Lupyan G, Bergen B. 2016. How Language Programs the Mind. *Top. Cogn. Sci.* 8(2):408–24

Majid A, Boster JS, Bowerman M. 2008. The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*. 109(2):235–50

Malt BC, Ameel E, Imai M, Gennari SP, Saji N, Majid A. 2014. Human locomotion in languages: Constraints on moving and meaning. *J. Mem. Lang.* 74:107–23

Malt BC, Gennari SP, Imai M, Ameel E, Saji N, Majid A. 2015. Where are the Concepts? What Words Can and Can’t Reveal. In *The Conceptual Mind: New Directions in the Study of Concepts*, eds. E Margolis, S Laurence. MIT Press

Malt BC, Gennari S, Imai M, Ameel E, Tsuda N, Majid A. 2008. Talking About Walking: Biomechanics and the Language of Locomotion. *Psychol. Sci.* 19(3):232–40

Manning CD. 2015. Computational Linguistics and Deep Learning. *Comput. Linguist.* 1–12

Manning CD, Raghavan P, Schütze H. 2008. *Introduction to Information Retrieval*, Vol. 1. Cambridge university press Cambridge

Manning CD, Schütze H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press

Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McCloskey D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. . 52:55–60

March JG. 1991. Exploration and Exploitation in Organizational Learning. *Organ. Sci.* 2(1):71–87

Markus HR, Kitayama S. 1991. Culture and the Self: Implications for Cognition, Emotion, and Motivation. *Psychol. Rev.* 98(2):224–53

Marshall EA. 2013. Defining population problems: Using topic models for cross-national comparison of disciplinary development. *Poetics*. 41(6):701–24

Massy WF. 1965. Principal Components Regression in Exploratory Statistical Research. *J. Am. Stat. Assoc.* 60(309):234–56

Mathieu J, Maynard MT, Rapp T, Gilson L. 2008. Team Effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse Into the Future. *J. Manag.* 34(3):410–76

Mcauliffe JD, Blei DM. 2008. Supervised topic models. *Adv. Neural Inf. Process. Syst.*, pp. 121–28

McFarland DA, Jurafsky Dan, Rawlings C. 2013a. Making the Connection: Social Bonding in Courtship Situations. *Am. J. Sociol.* 118(6):1596–1649

McFarland DA, Ramage D, Chuang J, Heer J, Manning CD, Jurafsky D. 2013b. Differentiating language usage through topic models. *Poetics*. 41(6):607–25

McGrath JE. 1984. *Groups: Interaction and Performance*. Prentice-Hall

Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, et al. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*. 331(6014):176–82

Mikolov T, Chen K, Corrado G, Dean J. 2013a. Efficient Estimation of Word Representations in Vector Space. *ArXiv13013781 Cs*

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013b. Distributed Representations of Words and Phrases and their Compositionality. *Adv. Neural Inf. Process. Syst. 26*, pp. 3111–3119. Curran Associates, Inc.

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013c. Distributed Representations of Words and Phrases and their Compositionality. *ArXiv13104546 Cs Stat*

Mikolov T, Yih W, Zweig G. 2013d. Linguistic regularities in continuous space word representations. , pp. 746–51. Association for Computational Linguistics

Miller IM. 2013. Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling approach. *Poetics*. 41(6):626–49

Miller JH, Page SE. 2009. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life: An Introduction to Computational Models of Social Life*. Princeton University Press

Mintzberg H, Raisinghani D, Théorêt A. 1976. The Structure of “Unstructured” Decision Processes. *Adm. Sci. Q.* 21(2):246–75

Mohr JW, Bogdanov P. 2013. Introduction—Topic models: What they are and why they matter. *Poetics*. 41(6):545–69

Mohr JW, Wagner-Pacifici R, Breiger RL, Bogdanov P. 2013. Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics. *Poetics*. 41(6):670–700

Moravcsik EA, ed. 2013. *Introducing Language Typology*. Cambridge University Press

Mosteller F, Wallace D. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley

Mukherjee S, Weikum G, Danescu-Niculescu-Mizil C. 2014. People on Drugs: Credibility of User Statements in Health Communities. *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 65–74. New York, NY, USA: ACM

Myerson RB. 2013. *GAME THEORY*. Harvard University Press

Nelson LK. 2015. *Political Logics as Cultural Memory: Cognitive Structures, Local Continuities, and Women's Organizations in Chicago and New York City*. Work. Pap.

Neumann J von, Morgenstern O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press

Nguyen TH, Shirai K. 2015. Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist.* http://www.anthology.aclweb.org/P/P15/P15-1131.pdf

Nickerson JA, Zenger TR. 2004. A Knowledge-Based Theory of the Firm—The Problem-Solving Perspective. *Organ. Sci.* 15(6):617–32

Niculae V, Kumar S, Boyd-Graber J, Danescu-Niculescu-Mizil C. 2015. Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game. *Proc. 53rd Annu. Meet. Assoc.*

Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Vol. 1 Long Pap., pp. 1650–1659. Beijing, China: Association for Computational Linguistics

Pachucki MA, Breiger RL. 2010. Cultural Holes: Beyond Relationality in Social Networks and Culture. *Annu. Rev. Sociol.* 36(1):205–24

Padgett J, Ansell C. 1992. Robust Action and the Rise of the Medici. *Am. J. Sociol.* 98:1259–1330

Padgett JF, Ansell CK. 1993. Robust Action and the Rise of the Medici, 1400-1434. *Am. J. Sociol.* 98(6):1259–1319

Pang B, Lee L. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* 2(1–2):1–135

Pelled LH, Eisenhardt KM, Xin KR. 1999. Exploring the Black Box: An Analysis of Work Group Diversity, Conflict, and Performance. *Adm. Sci. Q.* 44(1):1–28

Pellegrino F, Coupé C, Marsico E. 2011. A cross-language perspective on speech information rate. *Language*. 87(3):539–58

Pennebaker JW, Francis ME, Booth RJ. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway Lawrence Erlbaum Assoc.* 71:2001

Pennington J, Socher R, Manning C. 2014. Glove: Global Vectors for Word Representation. *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. EMNLP*, pp. 1532–1543. Doha, Qatar: Association for Computational Linguistics

Pentland A. 2012a. The new science of building great teams. *Harv. Bus. Rev.* 90(4):60–69

Pentland A. 2014. *Social Physics: How Good Ideas Spread-The Lessons from a New Science*. Penguin

Pentland A “Sandy.” 2012b. *Big Data’s Biggest Obstacles - HBR*. Harvard Business Review. https://hbr.org/2012/10/big-datas-biggest-obstacles

Pirolli PLT. 2007. *Information Foraging Theory: Adaptive Interaction with Information*. New York, NY, USA: Oxford University Press, Inc. 1st ed.

Prabhakaran V, Arora A, Rambow O. 2014a. Staying on topic: An indicator of power in political debates. *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. Doha Qatar Oct. Assoc. Comput. Linguist.* http://www.aclweb.org/old_anthology/D/D14/D14-1157.pdf

Prabhakaran V, Rambow O, Diab M. 2012. Predicting Overt Display of Power in Written Dialogs. *Proc. 2012 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, pp. 518–522. Stroudsburg, PA, USA: Association for Computational Linguistics

Prabhakaran V, Reid EE, Rambow O. 2014b. Gender and power: How gender and gender environment affect manifestations of power. *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. Doha Qatar Oct. Assoc. Comput. Linguist.* http://www1.cs.columbia.edu/~vinod/papers/EMNLP_genderpaper_final.pdf

Reagans R, McEvily B. 2003. Network Structure and Knowledge Transfer: The Effects of Cohesion and Range. *Adm. Sci. Q.* 48(2):240–67

Reagans R, Miron-Spektor E, Argote L. 2016. Knowledge Utilization, Coordination, and Team Performance. *Organ. Sci.* 27(5):1108–24

Redner S. 2001. *A Guide to First-Passage Processes*. Cambridge University Press

Ren Y, Argote L. 2011. Transactive Memory Systems 1985–2010: An Integrative Framework of Key Dimensions, Antecedents, and Consequences. *Acad. Manag. Ann.* 5(1):189–229

Resnik P, Garron A, Resnik R. 2013. Using topic modeling to improve prediction of neuroticism and depression. *Proc. 2013 Conf. Empir. Methods Nat.*, pp. 1348–1353. Association for Computational Linguistics$}$

Rickford JR, Duncan GJ, Gennetian LA, Gou RY, Greene R, et al. 2015. Neighborhood effects on use of African-American Vernacular English. *Proc. Natl. Acad. Sci.* 112(38):11817–22

Rivkin JW, Siggelkow N. 2003. Balancing Search and Stability: Interdependencies Among Elements of Organizational Design. *Manag. Sci.* 49(3):290–311

Rodriguez-Esteban R, Rzhetsky A. 2008. Six senses in the literature. The bleak sensory landscape of biomedical texts. *EMBO Rep.* 9(3):212–15

Rogers W. 1994. Regression standard errors in clustered samples. *Stata Tech. Bull.* 3(13):

Romero DM, Swaab RI, Uzzi B, Galinsky AD. 2015. Mimicry Is Presidential Linguistic Style Matching in Presidential Debates and Improved Polling Numbers. *Pers. Soc. Psychol. Bull.* 41(10):1311–19

Romney AK, Brewer DD, Batchelder WH. 1993. Predicting Clustering From Semantic Structure. *Psychol. Sci.* 4(1):28–34

Roscigno VJ, Hodson R. 2004. The Organizational and Social Foundations of Worker Resistance. *Am. Sociol. Rev.* 69(1):14–39

Rule A, Cointet J-P, Bearman PS. 2015. Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *Proc. Natl. Acad. Sci.* 201512221

Rzhetsky A, Foster JG, Foster IT, Evans JA. 2015. Choosing Experiments to Accelerate Collective Discovery. *Proc Natl Acad Sci U A*. 112(47):14569–74

Saavedra S, Hagerty K, Uzzi B. 2011. Synchronicity, instant messaging, and performance among financial traders. *Proc. Natl. Acad. Sci.* 108(13):5296–5301

Sacks H. 1995. Fall 1964-Spring 1965. In *Lectures on Conversation*, pp. 1–131. Wiley-Blackwell

Saji N, Imai M, Saalbach H, Zhang Y, Shu H, Okada H. 2011. Word learning does not end at fast-mapping: Evolution of verb meanings through reorganization of an entire semantic domain. *Cognition*. 118(1):45–61

Schank RC, Colby KM, Newell A. 1973. *Computer Models of Thought and Language*. WH Freeman San Francisco

Schegloff E. 1992. Introduction, Harvey Sacks' Lectures on Conversation. *Lect. Conversat.*

Schutt R, O'Neil C. 2013. *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media, Inc.

Schwarz W, Ischebeck A. 2003. On the relative speed account of number-size interference in comparative judgments of numerals. *J. Exp. Psychol. Hum. Percept. Perform.* 29(3):507–22

Schwenk CR. 1984. Cognitive simplification processes in strategic decision-making. *Strateg. Manag. J.* 5(2):111–28

Searle JR. 2010. *Making the Social World: The Structure of Human Civilization*. Oxford University Press

Shahaf D, Guestrin C, Horvitz E. 2012. Metro Maps of Science. *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1122–1130. New York, NY, USA: ACM

Shannon CE. 1948. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27(3):379–423

Shapin S. 1994. *A Social History of Truth : Civility and Science in Seventeenth-Century England*. Chicago: University of Chicago Press

Sheskin TJ. 1995. Computing mean first passage times for a Markov chain. *Int. J. Math. Educ. Sci. Technol.* 26(5):729–35

Shi F, Foster J, Evans J. 2014. *Weaving the Fabric of Science: Dynamic network models of science's unfolding structure*. Work. Pap., University of Chicago

Sim Y, Acree BD, Gross JH, Smith NA. 2013. Measuring ideological proportions in political speeches. *Proc. EMNLP*. http://hello.yanchuan.sg/assets/papers/sim2013measuring-slides.pdf

Simmons JP, Nelson LD, Simonsohn U. 2011a. False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* 0956797611417632

Simmons MP, Adamic LA, Adar E. 2011b. Memes Online: Extracted, Subtracted, Injected, and Recollected. *ICWSM*. 11:17–21

Simonton DK. 1999. Creativity as Blind Variation and Selective Retention: Is the Creative Process Darwinian? *Psychol. Inq.* 10(4):309–28

Smith A. 1759. *The Theory of Moral Sentiments*. London: A. Millar

Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *Proc. Conf. Empir. Methods Nat. Lang. Process. EMNLP*. 1631:1642

Spangler S, Wilkins AD, Bachman BJ, Nagarajan M, Dayaram T, et al. 2014. Automated Hypothesis Generation Based on Mining Scientific Literature. *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1877–1886. New York, NY, USA: ACM

Sridhar D, Foulds J, Huang B, Getoor L, Walker M. 2015. Joint models of disagreement and stance in online debate. *Annu. Meet. Assoc. Comput. Linguist. ACL*. http://www.anthology.aclweb.org/P/P15/P15-1012.pdf

Srikanth K, Harvey S, Peterson R. 2016. A Dynamic Perspective on Diverse Teams: Moving from the Dual-Process Model to a Dynamic Coordination-based Model of Diverse Team Performance. *Acad. Manag. Ann.* 10(1):453–93

Srivastava AN, Sahami M. 2009. *Text Mining: Classification, Clustering, and Applications*. CRC Press

Stein RT, Heller T. 1979. An empirical analysis of the correlations between leadership status and participation rates reported in the literature. *J. Pers. Soc. Psychol.* 37(11):1993–2002

Stewart GL, Courtright SH, Barrick MR. 2012. Peer-based control in self-managing teams: Linking rational and normative influence with individual and group performance. *J. Appl. Psychol.* 97(2):435–47

Stivers T, Enfield NJ, Brown P, Englert C, Hayashi M, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci.* 106(26):10587–92

Stone PJ, Kirsch J, Cambridge Computer Associates. 1966. *The General Inquirer : A Computer Approach to Content Analysis*. Cambridge, Mass.: M.I.T. Press

Stovel K, Shaw L. 2012. Brokerage. *Annu. Rev. Sociol.* 38(1):139–58

Stymne S, Hardmeier C, Tiedemann J, Nivre J. 2013. Feature weight optimization for discourse-level SMT. *DiscoMT Discourse Mach. Transl. 2013 9 August 2013 Sofia Bulg.*, pp. 60–69. Association for Computational Linguistics

Sudhof M, Goméz Emilsson A, Maas AL, Potts C. 2014. Sentiment Expression Conditioned by Affective Transitions and Social Forces. *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1136–1145. New York, NY, USA: ACM

Taddy M. 2013. Multinomial Inverse Regression for Text Analysis. *J. Am. Stat. Assoc.* 108(503):755–70

Tan C, Lee L, Pang B. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Vol. 1 Long Pap.*, pp. 175–185. Baltimore, Maryland: Association for Computational Linguistics

Tangherlini TR, Leonard P. 2013. Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics*. 41(6):725–49

Tausczik YR, Pennebaker JW. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29(1):24–54

Tetlock PC. 2007. Giving content to investor sentiment: The role of media in the stock market. *J. Finance*. 62(3):1139–68

Thierry G, Athanasopoulos P, Wiggett A, Dering B, Kuipers J-R. 2009. Unconscious effects of language-specific terminology on preattentive color perception. *Proc. Natl. Acad. Sci.* 106(11):4567–4570

Thompson LL. 2008. *Making the Team: A Guide for Managers*. Pearson/Prentice Hall

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 267–88

Tost LP, Gino F, Larrick RP. 2012. When Power Makes Others Speechless: The Negative Impact of Leader Power on Team Performance. *Acad. Manage. J.* 56(5):1465–86

Tsur O, Calacci D, Lazer D. 2015. Frame of mind: Using statistical models for detection of framing and agenda setting campaigns. *Proc ACL*. http://www.aclweb.org/anthology/P/P15/P15-1157.pdf

Tsytsarau M, Palpanas T, Castellanos M. 2014. Dynamics of News Events and Social Media Reaction. *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 901–910. New York, NY, USA: ACM

United Nations. 2016. *United Nations Data*. http://data.un.org/

Uzzi B. 1997. Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness. *Adm. Sci. Q.* 42(1):35–67

van Atteveldt WH. 2008. Semantic network analysis: Techniques for extracting, representing, and querying media content

Van de Vliert E. 2013. Climato-economic habitats support patterns of human needs, stresses, and freedoms. *Behav. Brain Sci.* 36(5):465–80

Viesturs E, Roberts D. 2007. *No Shortcuts to the Top: Climbing the World's 14 Highest Peaks*. Broadway Books

Vilhena D, Foster J, Rosvall M, West J, Evans J, Bergstrom C. 2014. Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication. *Sociol. Sci.* 1:221–38

Vuong QH. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*. 57(2):307–33

Wallace BC, Trikalinos TA, Laws MB, Wilson IB, Charniak E. 2013. A Generative Joint, Additive, Sequential Model of Topics and Speech Acts in Patient-Doctor Communication. *EMNLP*, pp. 1765–1775. http://www.aclweb.org/anthology/D13-1182.pdf

Wallach HM. 2006. Topic Modeling: Beyond Bag-of-words. *Proc. 23rd Int. Conf. Mach. Learn.*, pp. 977–984. New York, NY, USA: ACM

Wegner DM. 1987. Transactive Memory: A Contemporary Analysis of the Group Mind. In *Theories of Group Behavior*, pp. 185–208. Springer, New York, NY

Weick KE, Roberts KH. 1993. Collective Mind in Organizations: Heedful Interrelating on Flight Decks. *Adm. Sci. Q.* 38(3):357–81

Whissell C. 1989. The dictionary of affect in language. *Emot. Theory Res. Exp.* 4(113–131):94

Whorf BL, Carroll JB, Levinson SC, Lee P. 2012. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press

Willey MM. 1926. *The Country Newspaper: A Study of Socialization and Newspaper Content*. University of North Carolina Press

Winawer J, Witthoft N, Frank MC, Wu L, Wade AR, Boroditsky L. 2007a. Russian blues reveal effects of language on color discrimination. *Proc. Natl. Acad. Sci.* 104(19):7780–85

Winawer J, Witthoft N, Frank MC, Wu L, Wade AR, Boroditsky L. 2007b. Russian blues reveal effects of language on color discrimination. *Proc. Natl. Acad. Sci.* 104(19):7780–7785

Wittgenstein L. 2010. *Philosophical Investigations*. John Wiley & Sons

Woodward JL. 1934. Quantitative Newspaper Analysis as a Technique of Opinion Research. *Soc. Forces*. 12(4):526–37

Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW. 2010a. Evidence for a collective intelligence factor in the performance of human groups. *science*. 330(6004):686–88

Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW. 2010b. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science*. 330(6004):686–88

World Bank. 2016. *World Bank Open Data | Data*. https://data.worldbank.org/

Yang D, Wen M, Rose C. 2015. Weakly Supervised Role Identification in Teamwork Interactions. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Vol. 1 Long Pap.*, pp. 1671–1680. Beijing, China: Association for Computational Linguistics

Youn H, Sutton L, Smith E, Moore C, Wilkins JF, et al. 2016. On the universal structure of human lexical semantics. *Proc. Natl. Acad. Sci.* 113(7):1766–71

Yu B, Kaufmann S, Diermeier D. 2008. Classifying party affiliation from political speech. *J. Inf. Technol. Polit.* 5(1):33–48

Yu D, Tyshchuk Y, Ji H, Wallace W. 2015. Detecting Deceptive Groups Using Conversations and Network Analysis. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Vol. 1 Long Pap.*, pp. 857–866. Beijing, China: Association for Computational Linguistics