The Ontology of Communications:

Capturing Meaning from Organizational Communications

by

Gareth D. Keeves

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Business Administration)
in the University of Michigan
2018

Doctoral Committee:

Professor James D. Westphal, Chair
Professor Gerald F. Davis
Assistant Professor Derek Harmon
Professor Mark S. Mizruchi

Gareth D. Keeves

gkeeves@umich.edu

ORCID iD: 0000-0003-1379-0278

ACKNOWLEDGMENTS

Throughout my time in Michigan, I have been fortunate to receive the support and encouragement of so many friends and colleagues, which has allowed me to develop as a scholar, find my research passion, and has culminated in the writing of this dissertation.

First, I would like to thank my committee: Jim Westphal, Jerry Davis, Mark Mizruchi, and Derek Harmon, each of whom has had a significant impact on the direction, structure, and shape of my work. Since day one, Jim has provided foundational guidance, not only on research but also my broader academic outlook. From crafting ideas and developing an argument, through structuring and framing a paper, working with him has given me invaluable insight into the academic process. The perspective that he has ingrained will undoubtedly shape my entire academic career, providing the theoretical basis to build my subsequent work. In every way, Jim has been a fantastic mentor, always generous with his time, and there to provide support, guidance, and encouragement; there could not have been a better person for me to have worked with. My path to this dissertation was also significantly influenced by Jerry and the big-data camp that he organized in my first year. His enthusiasm for computational analysis helped me to see the potential of this rapidly evolving field within organizational research, and his encouragement and suggestions have helped broaden my thinking while helping me envisage the overall contribution of my work. I am also very grateful for the support that I have received from Mark throughout my Ph.D. The sociology class that I took with him early in the program fundamentally shaped my outlook of the field, helping me make connections between my research and sociological theories, and situate my work within the broader social science

literature. His thoughtful suggestions have helped me refine the framing and positioning of my work, sharpening my contribution to the overall development of social science theories. I am also very appreciative to have received the guidance of Derek. His insights into the impact of the structuring of language and deep understanding of the philosophies of meaning have allowed me to more clearly see the potential of research considering the multiple layers of structuring within textual information. Not only has Derek shaped my thinking on language, but he has also been extremely generous with help and encouragement, and his enthusiasm has reinforced the passion that I have for my research.

I have also benefited immensely from the overall Michigan community, the broad multidisciplinary training, and the acceptance of new approaches and ways of thinking that have given me the scholarly freedom to develop this dissertation. One of the most influential components of my training was the strategy classes completed in the formative stage of my time in the program; classes that shaped my trajectory, and still fundamentally influence my entire outlook on the strategy field. As I look back, my overall interest in understanding how communication evolves may be traced in part to Felipe Csaszar's Organizational Cognition class; my outlook on the evolution of communications parallels work that I was exposed to on organizational evolution. As I complete this dissertation and begin considering opportunities for extending the approach developed, possibilities to draw from the Carnegie research that I was introduced to in his class are increasingly resonating. Michael Jensen's Sociology of Strategy class also profoundly impacted my thinking, including my overall theoretical outlook and my appreciation of the need for a clear connection between theoretical constructs and how those constructs are measured. Indeed, this perspective helped me see limitations in existing approaches to characterize textual information and the opportunities from systematically

capturing a representation of what is said. Gautam Ahuja's Theoretical Perspectives in Strategy class also fundamentally shaped my entire understanding of the strategy field. His class continues to influence how I synthesize the key perspectives, helping me make connections between papers and identify discrepancies in the literature. Minyuan Zhao's International Business class also helped broaden my thinking, and as I consider opportunities to expand my approach to capture multi-lingual representations of text, the theories that we discussed in her class are again influencing my outlook. Finally, Brian Wu's Boundaries of the Firm class allowed me to appreciate the unique contribution and importance of economic perspectives in strategy. Overall, one of the key strengths of the strategy field comes from its diversity in perspectives, and the broad awareness and appreciation for the different areas of the field that I have gained throughout my doctoral training underpin my ability to identify connections between my work and other areas.

Of course, my training extends far beyond just the strategy classes that I have taken, and my thoughts are undoubtedly influenced by the interactions with scholars from across the business school and beyond. Various conversations that I have had with Guy Shani, Sun Hyun Park, Raji Kunapuli, and Ronnie Lee have helped me get where I am today. I was also very fortunate to have had the opportunity to collaborate with Suzan Ashford and Madeline Ong on a book chapter, through which I learned a great deal on the writing process and how to structure my ideas. My outlook is also influenced in distinct ways from the theories and perspectives that I was exposed to in other doctoral classes including Greta Krippner's Theories and Practices of Sociology class; Phoebe Ellsworth's Social Psychology class; David Mayer's Complex Cognitive Processes class; and Jeffrey Smith's Econometrics class. Jordan Siegel has also been a

fantastic source of academic guidance, and Brian Jones and the Doctoral office have provided fantastic support throughout the program.

My Michigan experience would not have been complete were it not for the wealth of friendships that I have made. I was so fortunate to have Casidhe Horan Troyer, Ashley Hardin, Reginald Edwards, Cassandra Chambers and Pete Aceves in my cohort – friendships that will each transcend my time at Michigan. I am also extremely grateful to my parents, who have always been there for me, providing help throughout the program. Finally, I have been so lucky to have met Kelsey, who has provided so much support throughout the journey. Overall, my time at Ann Arbor has been terrific, and I am so appreciative of everyone who has helped me develop and get where I am today.

PREFACE

One of the most profound pieces of advice that I was given during my time at Michigan was to envision the research landscape in 100 years time; to consider the questions that will have been examined, the insight that will have been gained, and the new dimensions that our theories will have explored and explained. While I cannot predict five years into the future, let alone one hundred, I can imagine a very different field than we currently operate in. Specifically, I can imagine a world where it is possible to examine the evolution of meaning, the structuring of that meaning, and the dynamics of how the structuring of meaning evolves. A future where it is feasible to capture a representation of everything that is said, how this aggregates, and how this aggregation evolves. In a similar way to how empirical analysis complemented early case studies, enriching the types of questions that could be analyzed, I see a world where the ability to systematically capture meaning allows us to push the boundaries of our understanding of the social world.

Since it is easy to ignore good advice, for nearly five years, it remained dormant, tucked away, but largely out of mind. However, the advice had a second component, and if I had paid more attention to that, maybe the length of my Ph.D. would have been much less. Specifically, I was advised that once distant research areas can be imagined, the next step is to implement them; to explore the questions that researchers of the future will ask, to develop the approaches to examine them, and to create the envisioned world. While I cannot re-write the past five years, I can create the next. This is my attempt to change theoretical discussion by opening new dimensions on which research can be based, helping to create the advancements that I believe are

not only feasible, but are an important component to the overall impact and vibrancy of the management and strategy field. It is an attempt to infuse a theoretical basis into large-scale textual analysis, enriching theory by enabling research questions that inherently require large volumes of rich, nuanced textual data.

Specifically, this dissertation develops and implements an approach to characterize textual information, transforming raw text into a consistent representation of what is said while preserving the structure. It is an approach to capture meaning, and the structuring of that meaning, en masse, such that it becomes feasible to envision and examine a new set of questions on how communications and discussions evolve. This dissertation is intended as the first step to create the world that I can imagine – the stepping-stone that illustrates the potential of synthesizing meaning while laying the foundations to make this prospect a reality. While it is impossible to predict the future of academia, through this dissertation, it is my intention to be firmly situated in the center of shaping it.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii
PREFACE vi
LIST OF TABLES x
LIST OF FIGURES xi
LIST OF APPENDICES xiv
ABSTRACT xvi

CHAPTER
I. Introduction and Motivation 1
General Introduction 1
Overview of Existing Computational Approaches 9
Unsupervised Computational Linguistic Approaches 9
Supervised Machine-Learned Classification of Texts 11
Information Extraction Approaches 13
Overview of the Approach Developed in this Dissertation 17
Overview of the Dissertation 23

II. Qualitative Development Process **24**
Unit of Analysis and Primary Information Types 24
Development of the Primary Ontologies 27
Removal of Surface Level Variation 32

III. Overview of the Information Extraction Process **34**
Stage 1: Concept Recognition 34
Stage 2: Semantic Interpretation 35
Stage 3: Standardization 36

IV. Managerial Backgrounds Implementation Specifics **38**
Extension of the Information Extraction Approaches to Represent Entire Sentences 39
Concept Recognition 42
Semantic Interpretation 43
Standardization and Verification of Concepts 45

V. Dimensions of Interest, Aggregation, and Comparisons **53**
Consideration of Theoretically Interesting Dimensions 53
Mathematical Approaches to Compare Text 56
Facilitation of Like-for-Like Comparisons 59
Structuring of Information 61

VI. Large-Scale Theoretically-Centered Textual Research Opportunities **63**
Individual Top Managers 63
Audience-Specific Impression Management 65
Conformance to Personas 66
Individual Top Managers: General Discussion 67
Overall Top Management Team 68
Presentation of the Diversity of the Top Management Team 69
Presentation of Board Control 71

VII. Discussion **72**
Connections Between the Approach and High-Level Logics 76
Development Process Going Forward 78

APPENDICES **82**

BIBLIOGRAPHY **136**

LIST OF TABLES

TABLE

Table II.1 Illustration of Sentences Typical of Each Information Type. 27
Table II.2 Summary of Surface-Level Variation 32
Table IV.1 Summary of the Standardization Approaches 47
Table IV.2 Summary of Validation Approaches 51
Table V.1 Mathematical Operations Feasible with Ontologies 58
Table V.2 Approaches for Extracting the Order 62
Table A.1 Summary of Text Colors Used in the Dissertation 83
Table C.1 Summary of the Primary Concepts 86
Table C.2 Summary of the Connecting Concepts 87
Table D.1 Overall Summary of the Sub-Ontologies 90
Table D.2 Summary of Date-Like Events 91
Table D.3 Summary of Qualification Levels 94
Table D.4 Summary of Qualification Subjects 94
Table D.5 Most Common Standardized Universities 99
Table D.6 Summary of Managerial Functional Areas 102
Table D.7 Primary Managerial Levels 104
Table D.8 Summary of Main Entity Types 106
Table D.9 Primary Committee Types 109
Table D.10 Summary of Stock Exchanges 113
Table D.11 Summary of Characterization Areas 116
Table D.12 Summary of Experience Types 123
Table D.13 Summary of Professional License Areas 127
Table E.1 Sources of Non-Conformance to Ontologies 129
Table E.2 Proportion and Number of Terms Classified by Sub-Ontology Type 130

LIST OF FIGURES

FIGURE

Figure I.1 Illustrating Raw Text Populated to an Ontology 14
Figure I.2 Example of a Sentence Discussing a Manager's Position History Represented in a Standardized Manner 19
Figure I.3 Example of a Sentence Discussing a Manager's Qualifications Represented in a Standardized Manner 20
Figure I.4 Example of a Sentence Discussing a Manager's Experiences Represented in a Standardized Manner 20
Figure I.5 Example of a Sentence Discussing a Manager's Professional License Represented in a Standardized Manner 20
Figure II.1 Illustration of Sentences Split into the Five Primary Ontologies 27
Figure II.2 Overall Sentence Representations 28
Figure II.3 Managerial Position in the POSITION Sub-Ontology 29
Figure II.4 Illustration of Managerial Titles Dissected by Key Dimensions 30
Figure II.5 Managerial Position Standardized in the POSITION, EXPERIENCE, QUALIFICATION, and PROFESSIONAL_LICENSE Sub-Ontologies 32
Figure II.6 Illustration of How Different Sentence Constructions Standardize 33
Figure III.1 Illustration of the Process by which Information is Populated to the Ontologies 36
Figure III.2 Illustration of Standardization of Concepts 37
Figure IV.1 Illustration of Indicative Sentence Complexity and the Intent of the Process 38
Figure IV.2 Illustration of How Increasing the Sentence Length Increases Complexity 40
Figure IV.3 Illustration of Grouping Concepts 41
Figure IV.4 Summary of the Overall Process to Classify Terms into Concepts. 42
Figure IV.5 Overall Illustration of Groupings 44
Figure IV.6 Example of Interpreting the Semantic Relationships for a Single Sentence 45
Figure IV.7 Illustration of Dissecting Concepts with Sub-Concepts 46
Figure IV.8 Illustration of Dissecting Sub-Concepts to Properties 46
Figure IV.9 Illustration of Standardization Through Dissection 48
Figure IV.10 Illustration of Verification Through Dissection 48
Figure IV.11 Standardization of EDUCATION_INSTITUTION and LOCATION Concepts 50
Figure V.1 Illustration of Position Comparisons 55
Figure V.2 Illustration of Position Comparisons Across Multiple Sentences 56
Figure V.3 Illustration of How the Process Integrates with Traditional Research 59
Figure V.4 Illustration of Comparison of Overall Top Management Team 61
Figure VI.1 Impression Management Cycle 68
Figure VI.2 Illustrations of the Multiple Layers of Structure 69
Figure VI.3. Stylized Illustration of Different Forms of Managerial Diversity. 70

Figure D.1 Specification of the DATE Sub-Ontology 92
Figure D.2 Examples of the DATE Sub-Ontology 92
Figure D.3 Specification of the LENGTH_OF_TIME Sub-Ontology 93
Figure D.4 Examples of the LENGTH_OF_TIME Sub-Ontology 94
Figure D.5 Specification of the DEGREE Sub-Ontology 97
Figure D.6 Examples of the DEGREE Sub-Ontology 98
Figure D.7 Example of a Split-Degree in the DEGREE Sub-Ontology 97
Figure D.8 Specification of the EDUCATION_INSTITUTION Sub-Ontology 98
Figure D.9 Examples of the EDUCATION_INSTITUITON Sub-Ontology 99
Figure D.10 Specification of the LOCATION Sub-Ontology 100
Figure D.11 Examples of the LOCATION Sub-Ontology 100
Figure D.12 Specification of the FUNCTIONAL_AREA Sub-Ontology 101
Figure D.13 Examples of the FUNCTIONAL_AREA Sub-Ontology 102
Figure D.14 Examples of acronyms in the FUNCTIONAL_AREA Sub-Ontology 102
Figure D.15 Specification of the MANAGEMENT_LEVEL Sub-Ontology 103
Figure D.16 Examples of the MANAGEMENT_LEVEL Sub-Ontology 105
Figure D.17 Specification of the COMPANY-INDUSTRY_AREA Sub-Ontology 106
Figure D.18 Examples of the COMPANY-INDUSTRY_AREA Sub-Ontology 107
Figure D.19 Specification of the MANAGEMENT_TITLE Sub-Ontology 108
Figure D.20 Examples of the MANAGEMENT_TITLE Sub-Ontology 108
Figure D.21 Specification of the COMMITTEE Sub-Ontology 109
Figure D.22 Examples of the COMMITTEE Sub-Ontology 110
Figure D.23 Specification of the BOARD Sub-Ontology 111
Figure D.24 Examples of the BOARD Sub-Ontology 111
Figure D.25 Specification of the FIRM_FINANCIAL Sub-Ontology 112
Figure D.26 Examples of the FIRM_FINANCIAL Sub-Ontology 112
Figure D.27 Specification of the LISTING-OWNERSHIP Sub-Ontology 113
Figure D.28 Examples of the LISTING-OWNERSHIP Sub-Ontology 114
Figure D.29 Specification of the CHARACTERIZATION Sub-Ontology 115
Figure D.30 Examples of the CHARACTERIZATION Sub-Ontology 116
Figure D.31 Specification of the COMPANY_DESCRIPTION Sub-Ontology 117
Figure D.32 Examples of the COMPANY_DESCRIPTION Sub-Ontology 117
Figure D.33 Specification of the COMPANY_NAME Sub-Ontology 118
Figure D.34 Examples of the COMPANY_NAME Sub-Ontology 118
Figure D.35 Specification of the PERSON_NAME Sub-Ontology 119
Figure D.36 Examples of the PERSON_NAME Sub-Ontology 119
Figure D.37 Specification of the BACKGROUND_DETAILS Primary-Ontology 120
Figure D.38 Examples of the BACKGROUND_DETAILS Primary-Ontology 120
Figure D.39 Specification of the POSITION Sub-Ontology 121
Figure D.40 Examples of the POSITION Sub-Ontology 121
Figure D.41 Examples of the POSITION Sub-Ontology with Boards/Committees 122
Figure D.42 Examples of the POSITIONS Primary-Ontology 122
Figure D.43 Specification of the EXPERIENCE Sub-Ontology 123

Figure D.44 Examples of the EXPERIENCE Sub-Ontology 123
Figure D.45 Specification of the QUALIFICATION Sub-Ontology 124
Figure D.46 Examples of the QUALIFICATION Sub-Ontology 124
Figure D.47 Examples of the QUALIFICATIONS Primary-Ontology 125
Figure D.48 Specification of the PROFESSIONAL_LICENSE Sub-Ontology 126
Figure D.49 Examples of the PROFESSIONAL_LICENSE Sub-Ontology 127
Figure E.1 Cumulative Distribution of Managerial Titles (Logarithmic Scale) 128
Figure E.2 Illustration of Connecting Company Areas to Associated NAICS codes through Wikipedia Data 132
Figure F.1 Mocked-Up of Ontologies for Selection Decisions 134
Figure F.2 Illustrations of Ontologies Populated in Spanish, with either a) English or b) Spanish Ontology Labels 135

LIST OF APPENDICES

APPENDIX

A. Text Colors Used in Dissertation 83

B. Collection of Managerial Background Data 84

C. Classifying the Text to Concepts 85
Primary concepts 85
Connecting concepts 86

D. Specifications of the Ontologies 91
Dates and Date-Like-Events 91
Length of Time 93
Degree 94
Education Institution 98
Location 100
Functional Area 101
Management Level 103
Company-Industry Area 106
Management Title 108
Committee 109
Board 111
Firm Financials 110
Listing-Ownership 111
Characterizations 113
Company Description 117
Company Name 118
Person Name 119
Background Details 120
Position 121
Experience 123
Qualification 124
Professional License 126

E. Standardization and Verification 128
Summary of Standardization and Verification Status 128
Approaches Under Development to Classify and Verify Fuzzy Concepts 130
Initial Development of Automated Standardization of Industry and Company Areas to NAICS Codes 131

F. Expansion of the Approach 133

Expansion to Capture Top Manager Selection Decisions 133
Multi-Lingual Representations 134

ABSTRACT

This dissertation seeks to further scholarly understanding of the content and structure of organizational communications by developing an analytical approach to systematically transform communications into consistent ontologies. By representing the meaning of what is said in a consistent manner, while reflecting the underlying structuring of that information, this dissertation is intended to facilitate large-scale analysis of the evolution of meaning and different layers of structuring. Specifically, once textual information is transformed into consistent ontologies, it becomes feasible to examine the content of what is said, how that content is structured, and how sub-structures combine to overall meta-structures. As such, the overall approach developed is intended to enrich strategy, management, and social science research by allowing the development and testing of theories that inherently require large volumes of rich, nuanced data, such as the process by which high-level structuring of information evolves.

While the approach developed is general, able to be expanded across the social sciences, this dissertation focuses on capturing and representing meaning from managerial backgrounds. Specific consideration is given to illustrate how by removing surface-level variations, such as acronyms, synonyms, and superficial differences in sentence construction, inconsistently written sentences can be transformed to consistent ontologies. This dissertation also illustrates how consistently representing the key dimensions of the experiences, positions, qualifications and professional licenses discussed in managerial backgrounds, provides the basis for capturing theoretically meaningful concepts, that can be measured across the population of managers. By providing a path by which theoretically motivated constructs can be developed and utilized, this

dissertation is intended to bridge theoretically orientated social science research with advancements in data science, helping to facilitate the growth of theoretically-centered textual analysis, which has broad possibilities to enrich strategy and management theories.

CHAPTER I

Introduction and Motivation

General Introduction

There is no question that textual information represents a rich source of historical information to understand an array of firm actions. From analyzing strategic actions taken by organizations (Tripsas and Gavetti, 2000; Benner, 2010), to studying the dynamics of firm and stakeholder interactions (Chen and Hambrick, 1995; King, 2008), to understanding the overall evolution of a field or society (Fairclough, 1992; Barley and Kunda, 1992), the qualitative textual information produced by organizations, information intermediaries, and other stakeholders presents a wealth of detailed historical information for organizational researchers. Moreover, in addition to conveying factual information, communication provides opportunities to shape reality (Berger and Luckmann, 1967; Potter, 1996), and organizational communication is a way of managing audience perceptions (Bettman and Weitz, 1983; Elsbach, 2006; Fiss and Zajac, 2006). With communications underpinning and documenting organizational behaviors and how perceptions are shaped, it is unsurprising that scholars from a wide array of theoretical orientations have drawn on textual data to capture concepts of interest (e.g., Yoon and Park, 2004; Kennedy, 2005; Chatterjee and Hambrick, 2007), and that there is considerable scholarly interest in understanding organizational communication itself (e.g., Boje, 1991; Elsbach, 2006; Sillince, Jarzabkowski, and Shaw, 2012; Kahl and Grodal, 2016).

However, while communication underpins a substantial proportion of organizational theory and theoretical constructs, and a wealth of textual information is now easily available for researchers (including company filings, websites, conference calls and patents), the ability of researchers to systematically extract meaning from this information, or construct variables that closely map to concepts of theoretical interest, is limited. While qualitative studies give significant consideration to the content and form of firm communications (e.g., Martin et al., 1983; Elsbach, 1994; Fiol, 2002), more quantitative approaches often reduce complex, multifaceted theoretical concepts, to a list of keywords, the frequency of which are then counted. The list of constructs that researchers have attempted to capture via word counts is long, including the valence of the text (Loughran and McDonald, 2011; Bednar, Boivie, and Prince, 2013), future vs. past orientation (König et al., 2018), the 'grammar' of decision making (Crilly, Hansen, and Zollo, 2016) and institutional logics (Ocasio and Joseph, 2005). Although a particular dimension of meaning may be captured via vocabularies (Tausczik and Pennebaker, 2010; Loewenstein, Ocasio, and Jones, 2012), word usage is nevertheless just one lens through which language can be characterized (Fairclough, 1992), and the approach quickly becomes unfeasible for understanding complex ideas or how these complex ideas evolve over time. Moreover, although there is growing computational linguistics research on ways in which textual communications can be analyzed (e.g., Chowdhury, 2003; Aggarwal and Zhai, 2012), and certain fields such as biology and medicine have seen considerable interest in standardizing and extracting textual information (Cohen and Hersh, 2005; Simpson and Demner-Fushman, 2012; Lacey et al., 2017), there has been very little consideration of the ways in which computational linguistics can yield new theoretical insight in strategy and management research, nor has there

been a cumulative effort to systematically capture the spectrum of relationships discussed by firms in their communications.

This research endeavor seeks to enable analysis of the relationships and meaning discussed in organizational communications, by developing an approach to systematically capture and represent the nature and form of the meaning conveyed, which can ultimately be extended to the communications of the various stakeholders with which the organization interacts. Specifically, this research seeks to address three fundamental limitations with approaches that count words (or predefined phrases) that restrict theoretical development by constraining the types of concepts that can be measured; limitations that, as discussed later, are shared with other approaches which take words as the fundamental unit of analysis (e.g., topic modeling: Wilson and Joseph, 2015; Kaplan and Vakili, 2015; Bao and Datta, 2014).

The first limitation of word-based measures derives from the key assumption that meaningful concepts can be captured by analyzing words, or clusters of words, in isolation. Although this assumption may hold for certain concepts (e.g., arguably, the valence of the text Tausczik and Pennebaker, 2010; or high-level logics: Ocasio and Joseph, 2005), sentences are used to express relationships between concepts. At the simplest level, these relationships express connections between objects or attributes, for example, how much experience a particular manager has in a particular industry. Textual information can also express more nuanced and complicated relationships, such as justification for particular decisions, expectations of future states, and caveats to an argument. Since all but the simplest of relationships are comprised of multiple words, it is essentially impossible to characterize the underlying message, or meaning of the communications, by analyzing words in isolation.[1] Moreover, while it could be argued that

[1] While it could be argued that certain 'ideas' can be captured through a single word, such as '**success**' or '**pleased**', the meaning of even very simple concepts changes depending on the surrounding words: '**our**

word-count approaches could be extended to capture more complex concepts by searching for phrases (e.g., '**experience in the manufacturing sector**'), it quickly becomes unfeasible to foresee every way in which even the simplest relationships can be conveyed. Thus, while it may be possible to capture high-level themes or changes to those themes over time (Ocasio and Joseph, 2005), characterizing the relationships between concepts, and how the discussed relationships are evolving and diffusing, is essentially impossible through word counts alone. This may be especially true for capturing the relatively subtle, nuanced, and complex concepts in management and strategy theory. Thus, although textual archives may provide the best record of the process by which organizational, institutional, and societal change occurs (e.g., Maguire and Hardy, 2009; Funk and Hirschman, 2014), it is fundamentally difficult to systematically capture this change process for more quantitative analysis, with scaling limitations restricting ability to manually hand code meaning at the overall field level.

The second fundamental issue with current word-based approaches is that they give little consideration to structure; word counts ignore the syntax of how the words are combined to construct meaning (e.g., Matthews and Matthews, 1981; Van Valin and LaPolla, 1997), and how the components of a message are aggregated into an overall meta-structure.[2] Indeed, given the limitations of existing approaches to capture a representation of the relationships discussed in the

---

customers are pleased' and '**our shareholders are pleased**' convey different relationships. While it has been argued that meaning can be captured through topic modeling (Kaplan and Vakili, 2015) (an approach which still takes the word as the fundamental unit of analysis, but clusters documents together based on usage of sets of co-occurring words, or so-called 'topics') this ignores how slight differences in the words and syntax can result in substantial differences in meaning, and that documents with broadly similar themes can nevertheless express very different or opposing ideas. While topic modeling and related approaches may thus be suitable to group documents based on similar broad topics, the approaches were never designed to extract the specific relationships discussed in the text. Moreover, as discussed in more detail later, approaches that take the whole sentence as a unit of analysis (such as classifying sentences into a predefined categories through machine learning: e.g., Pang, Lee, and Vaithyanathan, 2002; Caruana and Niculescu-Mizil, 2006), are likewise ill-suited to identifying the relationships within sentences nor the structure of those sentences.

[2] The term meta-structures is used to expand upon the concept of 'narratives' (e.g., Franzosi, 1998; Martens, Jennings, and Jennings, 2007), a term that although also concerned with the overall structuring of a document, is used in a slightly narrower sense, to refer to the sequencing of the components of an event.

text, it is difficult to conceive how it could be possible to examine an added layer of complexity, namely how those underlying relationships are structured. Thus, while structure can be conceived of in many different ways (Fairclough, 1992), automated approaches focus on characterizing the language used (such as measures of complexity derived from sentence and word length: Courtis, 1986; Li, 2008; Rennekamp, 2012), rather than how the underlying components of the message are themselves constructed and presented. Specifically, there is very little consideration of the argumentation structure (Harmon, Green, and Goodnight, 2015), or how the order in which information is presented shapes interpretation (Leung, 2014). Research that does consider the impact of the structure of the language tends to be either qualitative in nature (e.g., Martin et al., 1983; Boje, 1991), or derive from hand-coding the structure of the sentences (e.g., Bettman and Weitz, 1983; Harmon, 2018).

While qualitative approaches, including hand coding of sentence structure, may yield substantial insight, their application is constrained to instances where it is feasible to read or hand code each sentence. Although it is possible to dismiss the scalability difficulties of qualitative approaches (e.g., that it is unnecessary to scale, or that the issue can simply be solved by scaling the number of research assistants), scalability difficulties pose a substantial constraint to theoretical development. This may be best illustrated by considering the next level of structure: how individual relationships are combined into overall meta-structures. While there are qualitative studies that examine how components of a message are combined into overall meta-structures (such as work on narratives or storytelling: Martin et al., 1983; Boje, 1991), these are limited to case studies or very small numbers of firms; there has been very little theoretical or empirical consideration of the causes of variations in meta-structures at the field level. Moreover, it is especially difficult to examine interactions between the meta-structures of different firms;

development and testing of theory examining how field-level meta-structures form and evolve is beyond the ability of even the largest conceivable number of research assistants to hand code. The lack of an ability to systematically 'capture' relationships, the structure of those relationships, and how those relationships are structured into overall meta-structures, thus explains the void of research beyond limited qualitative studies examining how structures and meta-structures form and evolve at the field level.

The final issue with using word-lists to capture constructs of interest is that the approach relies on the ability of researchers to specify lists of all relevant words. While this may be appropriate for capturing concepts that can be represented through a small number of words (or phrases), there are some concepts where it would essentially be impossible to anticipate all words (or phrases) in advance. Although theory may help define concepts of interest, theory typically offers little guidance of the specific words that underlie those concepts. For example, it would be unfeasible to define in advance all of the experiences that a manager may have (e.g., to compare a how a manager's prior experiences are recharacterized over time); while certain words are common and easy to anticipate (e.g., '`financial experience`'), other phrases are much more obscure and hard to anticipate (e.g., '`substantial accounting experience, with a focus on mergers and acquisitions`').[3] Similarly, while it would be possible to identify '`manufacturing experience`' and '`accounting experiences`' as types of experience, it is hard

[3] Multi-word offerings, in particular, are hard to identify because it is not possible to simply examine all of the individual words used. While, as discussed later, there are lots of entity-extraction tools that can extract certain types of information from text, without the need to specify words in advance (e.g., NLTK: Bird and Loper, 2004; or Stanford NER: Finkel, Grenager, and Manning, 2005), the information types that these tools capture is relatively limited to a narrow range of concepts, typically: company names, location names, people names, dates, and contact information; although there may be certain research questions that can be answered with just this type of information (e.g., using co-citations of company names to capture categories: Kennedy, 2005), simply extracting these types of information only captures a relatively narrow dimension of what is said, and certainly not the relationships between concepts.

to foresee all possible aggregations, such as '**manufacturing, accounting, and human resource experience**'.[4]

This dissertation tackles head-on the issue that current text analysis approaches are more concerned with counting 'words' than capturing the meaning and form of the communications, by developing an approach, initially targeted at strategy and management scholars, to extract and standardize the relationships conveyed in organizational communications, as well as the form and structure of those ideas, with the ultimate intent that this analysis approach can be extended across the social sciences. This research has a dual intention. First, by allowing scholars to capture more nuanced and complex theoretical constructs, it seeks to facilitate theoretical development. Since a close match between theory and measurement of theoretical constructs is generally required (Campbell and Stanley, 1963; Allen and Yen, 1979; Blackstone, 2012), theoretical nuance tends to reflect the capability of measuring that nuance in the underlying data. Limitations in being able to capture rich constructs may, in turn, lead to theoretical simplification. By allowing multi-faceted ideas to be directly extracted and easily manipulated, more rich and nuanced theory can be developed and tested. Moreover, in addition to being able to enrich theory by more closely capturing multi-faceted constructs, the ability to directly measure variables of interest (rather than distant proxies) allows theory on cross-medium communications to be more easily developed and tested. Specifically, being able to directly

[4] There are clearly other challenges that analyzing languages entails, for example, that the meaning of words can change depending on their usage, and that often just one word can fundamentally change the meaning of a sentence (Hannigan, 2015). This challenge becomes more apparent when moving past managerial backgrounds; for example, 'the risk of default is high' is clearly different to 'the risk of default is low', yet simply counting the 'risk' words ignores such difference.

capture nuanced concepts from large volumes of information facilitate comparisons despite the presence of substantial extraneous variation between mediums.[5]

The second intention of this research is that once it is possible to systematically capture nuanced information from text, it becomes possible to conceive of and analyze a whole new set of theoretical questions that inherently require large volumes of rich, nuanced data. For example, it is hard to envision how progress can be made in understanding field level evolution of the meta-structure of texts, without first making progress in systematically capturing and representing the meaning of the communications, and the aggregation of meaning into overall structures. Indeed, since many forms of communication are long in length (e.g., the text detailing managerial backgrounds is typically several pages per firm, and other components of company filings and conference calls can each take several hours to read), quantitative analysis of the structure of such documents, beyond measures of complexity derived from sentence and word length (e.g., Courtis, 1986; Li, 2008; Rennekamp, 2012), is inherently unfeasible for all but the smallest number of firms. Systematically capturing the relationships discussed in organizational communications not only facilitates quantitative studies on the structure of text, but it also makes it easy to expand this research across industries and time periods, allowing the contingencies under which relationships hold or are especially strong to be theorized and assessed.

[5] For example, as discussed in more detail later, while there is growing interest in audience-specific perceptions (Bourdieu, 1984; Jensen, Kim, and Kim, 2012; Ertug et al., 2016), there is little theory considering how firms tailor their communications to different audiences, or their ability to create audience-specific perceptions. A potential difficulty with analyzing this question is that communications using different mediums typically differ in ways that are unrelated to the variation that a specific research question seeks to examine. While similarity measures based on word-overlap may identify differences between mediums, other extraneous differences between the communication mediums may nevertheless make it hard to know what is driving the differences (e.g., whether it is just different words used to convey similar information, or whether actually different information is being presented). More targeted measures, specifically capturing the construct of interest, allow comparisons to be made across communication mediums, by specifically identifying the difference of interest (i.e., without capturing substantial, or nonsystematic, noise from other forms of variation).

Overview of Existing Computational Approaches

Before the approach developed in this paper is introduced, and the specific contributions to the management and strategy literatures are discussed, it is first useful to briefly summarize computational linguistic approaches. While computational linguistics is a broad field, and it is thus unfeasible to cover all aspects of the literature in detail here (for a recent overview, see: Jurafsky and Martin, 2018), key paradigms include: i) unsupervised (i.e. fully automated) characterizations; ii) machine-learned classification of texts into predefined categories; and iii) extraction of information into standardized ontologies. As discussed further below, while a lack of connections to theoretical constructs limits the general applicability of unsupervised approaches to test management and strategy theories, and machine-learned classification of sentences has limited ability to analyze the relationships conveyed within the sentences, information extraction offers opportunities for developing constructs that closely correspond to variables of theoretical interest to management and strategy scholars, as well as other researchers in the broader social sciences.

Unsupervised Computational Linguistic Approaches

There are a broad array of unsupervised computational linguistic approaches that characterize textual communications, without requiring any domain-specific understanding of the underlying material. Although topic models are one of the more widely used approaches in management and strategy research to date (e.g., Magerman, Looy, and Song, 2010; Wilson and Joseph, 2015; Kaplan and Vakili, 2015; Bao and Datta, 2014),[6] other uses of unsupervised approaches include data visualization (Grimmer and King, 2011) and determining document

[6] Topic modeling is a general term used to describe a variety of closely related latent clustering approaches, some of the most common being: Latent Semantic Analysis (LSA) (Papadimitriou et al., 2000); Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999); and Latent Dirichlet Allocation (LDA) (Pritchard, Stephens, and Donnelly, 2000; Blei, Ng, and Jordan, 2003).

keywords (Hulth, 2003; Matsuo and Ishizuka, 2004; Ercan and Cicekli, 2007). Each of the approaches builds on varying forms of basic linguistics principles, for example, that words occurring in similar contexts tend to be semantically close (the distributional hypothesis: Firth, 1957), or that pairs of co-occurring words tend to have similar semantic relations (the latent relation hypothesis: Turney, 2008). These basic principles enable text to be reduced to a vector, and inference between texts to be made using vector-based similarity measures (for a review see: Turney and Pantel, 2010).[7] The key advantage of such unsupervised approaches is their versatility; since it is unfeasible to conceive of every possible concept that may be discussed across the universe of all communications, basic linguistic assumptions allow documents to be compared without the need for *a priori* theory to inform relevant similarity dimensions (Turney and Pantel, 2010).

For many real-world applications, the use of unsupervised approaches to identify latent dimensions, without the need for *a priori* theory, is an advantage; it is often desirable that measures function well 'out of the box' without the need to determine, or extract, a basis for similarity. As such, implementations of unsupervised computational linguistic approaches are common, with applications including: suggesting related articles on news websites (Kanhabua, Blanco, and Matthews, 2011; BBC News Labs, 2016); analyzing the content of websites to determine contextually relevant adverts (Google, 2003); and identifying keywords from scholarly articles (Rose et al., 2010). The primary requirements in each of these settings is that the approach works in an automated manner, across the universe of encountered texts, and

[7] Since language has an unbounded number of possible words, that can, in turn, be organized in an unbounded number of ways, unsupervised approaches reduce the dimensions of the text (with different approaches based on different basic linguistic principles) (Turney and Pantel, 2010). Once the text is represented as a vector, it is easy to manipulate, and calculate similarity scores between documents. For example, topic modeling involves first creating buckets (or 'topics') of co-occurring words (Papadimitriou et al., 2000), (which relies on the distributional hypothesis, that words that are used in similar contexts are likely to have similar meanings: Firth, 1957); once topics have been created, a document can be represented as a vector based on usage of each of the different topics, and in turn, similarity scores (e.g., cosign similarity) calculated.

produces results that are deemed accurate (or sufficiently accurate) by humans; the specific basis, or assumptions, by which these approaches function is a secondary, (or non) consideration.

However, while it is undeniable that unsupervised approaches have substantial real-world applications, it is less clear that such approaches are well suited for developing and testing social science theories. It is generally accepted that theory testing requires a clear rationale for measuring predefined concepts of interest (e.g., Campbell and Stanley, 1963; Allen and Yen, 1979; Blackstone, 2012). While manual coding allows researchers to construct variables that closely match theoretical concepts, unsupervised computational linguistic approaches are explicitly not intended to capture predefined concepts. Moreover, since the predominant paradigm in quantitative management and strategy research (as well as much of the broader social sciences), is testing predefined theory, it is unlikely that inductive research (e.g., Magerman, Looy, and Song, 2010; Grimmer and King, 2011), where insight is drawn from the data in the absence of predefined theory, will have a substantial impact on the overall field.[8]

Supervised Machine-Learned Classification of Texts

An intermediary approach, which may have applications in testing management and strategy research questions (although for substantially different questions than this dissertation seeks to allow) involves training a machine learning model to automatically classify blocks of

[8] There may be a role for unsupervised approaches including identifying potentially relevant dimensions that the literature may not have previously considered (e.g., Bao and Datta, 2014), in a comparable manner to how qualitative research contributes to informing basis for theoretical consideration (e.g., Eisenhardt, 1989). However, while one of the arguments often presented in favor of latent approaches is that they do not impose assumptions or structure on the data, this is arguably an over-simplification. Specifically, latent approaches have their own sets of assumptions (e.g., that words can be interpreted in isolation: Turney and Pantel, 2010), and as such, the latent concepts still reflect some underlying assumptions of the analysis approach. Moreover, at least in social science research, it is typical to re-impose structure onto the data, if only to make sense of the latent concepts. For example, topic modeling gives no direct guidance as to the underlying meaning behind topics: such interpretation typically requires researchers to deduce the likely meaning of the buckets of words (e.g., Bao and Datta, 2014: 1382). As such, even if word lists are derived without direct manual involvement, inferring the meaning by labeling the topics re-imposes structure, inherently relying on human knowledge of likely corresponding concepts, and in turn further offsetting one of the purported benefits of latent approaches.

text, such as sentences or whole documents, into predefined categories (Pang, Lee, and Vaithyanathan, 2002; Caruana and Niculescu-Mizil, 2006). This approach has wide commercial applications, for example using manually flagged 'inappropriate' or spam communications to train pattern recognition algorithms to automatically identify other material that shares similar characteristics (Drucker, Wu, and Vapnik, 1999; Sebastiani, 2002). In the social sciences, researchers could, for example, manually classify a relatively small number of sentences or documents into predefined categories, and then use machine learning to automatically classify a much larger number of sentences or documents into these predefined categories. A benefit of this approach for management and strategy research is its ability to categorize material into predefined concepts desired by researchers, rather than 'latent' concepts in unsupervised approaches.

However, while this may be appropriate for some research questions, analyzing sentences (or documents) as the fundamental unit of analysis introduces its own limitation: namely, it is hard to reduce meaning to a small number of categories. For example, a simple sentence may be: '`We use hedging to minimize foreign exchange risk`'. While it is possible to classify this sentence as 'hedging' or 'foreign exchange risk', and this may be suitable for certain purposes, it is less suited to capturing the relationships underlying the sentence, specifically that the use of hedging minimizes the firm's foreign exchange risk.[9] This limitation restricts the sort of questions that can be answered through machine-learned classification of texts, to the same sort of questions that it is possible to investigate with word counts. While overall trends such as 'high-level institutional logics' could be captured by classifying sentences into concepts, this approach is less suited to capturing the specific relationships, or the micro-foundations

[9] As discussed further in the Discussion Section, the approach developed in this dissertation is intended to be applicable beyond managerial backgrounds, to capture justifications such as this.

underpinning institutional logics (e.g., Seo and Creed, 2002; Munir and Phillips, 2005; Lawrence and Suddaby, 2006), or to analyze how these foundations are adapted and evolve over time. Thus, although machine-learned classification of text may be more appropriate for testing some types of management and strategy research questions than unsupervised approaches, it is nevertheless inherently limited to situations where the research question can be examined by classifying units of texts (e.g., sentences or documents) into a limited number of predefined categories.

Information Extraction Approaches

The final group of approaches, which this paper draws upon and explains in greater detail in due course, involves directly extracting information from text and representing this information in a standardized format, or ontology (Aggarwal and Zhai, 2012; Biega, Kuzey, and Suchanek, 2013). While most factual information throughout history was originally written as text (e.g., books, research papers, and more recently webpages), there are many benefits in representing the information in a consistent manner. Once in an ontology, it is possible to easily query the information, make comparisons between items, aggregate the information to produce an overall summary, and to connect the information to other databases (Auer et al., 2007; Suchanek, Kasneci, and Weikum, 2007).[10] The intent of the process is outlined in Figure I.1.

[10] While the terms 'unstructured data' is used in computer science to refer to raw text, and 'structured data' is used to refer to data that is represented into a standardized format or ontology (e.g., Gupta and Lehal, 2009), these terms are avoided to prevent confusion with discussion of linguistic structure discussed in this paper (e.g., Fairclough, 1992).

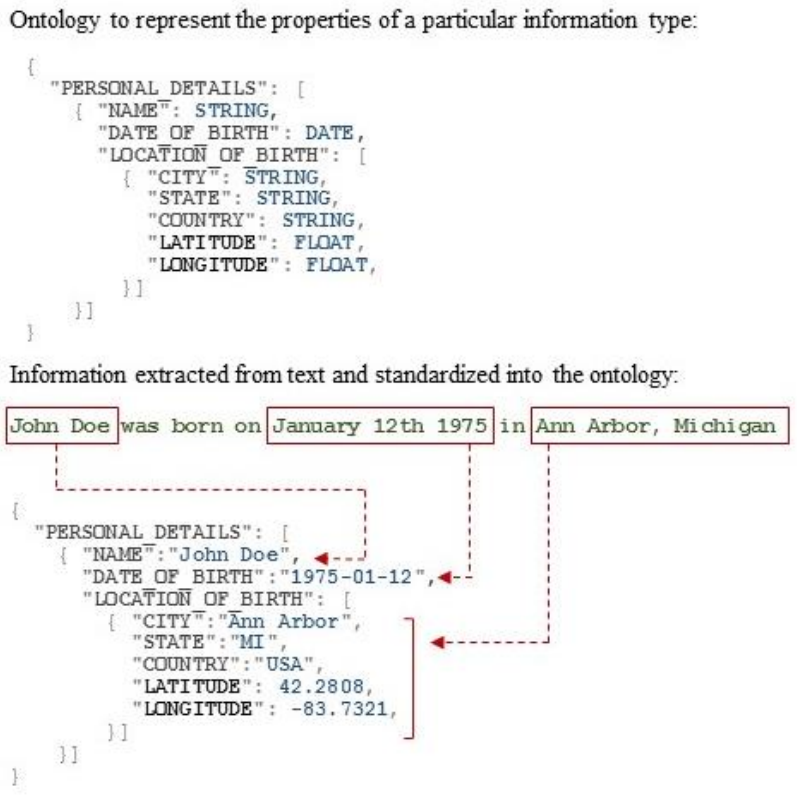


Figure I.1. Illustrating Raw Text Populated to an Ontology

Although information extraction is an area of continuing research, and specific implementations vary, the overall approaches are broadly similar (e.g., Suchanek, Kasneci, and Weikum, 2007; Gupta and Lehal, 2009; Han, Pei, and Kamber, 2012). Specifically, the approach involves first identifying concepts such as people, companies, and dates,[11] and then inferring meaning from the semantic-relationships (or connecting words) between the concepts (Chklovski and Pantel, 2004; Lample et al., 2016). For example, in the sentence '**John Doe was born on January 12th 1975 in Ann Arbor, Michigan**' in Figure I.1, an entity recognition model could

[11] While the terms including 'entity recognition' and 'entity extraction' are commonly used in the computer science literature to describe the process of recognizing concepts in text (e.g., Lample et al., 2016: 1), reflecting that they are typically used to capture entities such as people, companies and places, the term 'concept recognition' is used throughout this dissertation to reflect the broadening of the information types identified in the text, to include more abstract concepts such as experiences and company descriptions.

be trained using a sample of tagged sentences, to recognize '`John Doe`' is likely a `PERSON_NAME`, and '`January 12th, 1975`' is likely a `DATE`. The connecting words '`was born on`', (as well as any synonymous connecting words as appropriate), could then be used to infer that a `PERSON_NAME`, called John Doe, was born on the `DATE` January 12th, 1975. Similarly, the approach can be extended to recognize that a date of birth, followed by the word '`in`' and then a `LOCATION`, indicates the birth location.[12]

There are a large number of active research projects that use similar information extraction approaches to harvest factual information from Wikipedia and the wider internet, representing the extracted information in a consistent, standardized format. These include: YAGO (Biega, Kuzey, and Suchanek, 2013); NELL (Carlson et al., 2010); PROSPERA (Nakashole, Theobald, and Weikum, 2011); SOFIE (Suchanek, Sozio, and Weikum, 2009) and DBpedia (Auer et al., 2007). Moreover, many commercial entities are also engaged in information extraction. For example, Google's Knowledge Graph initiative gathers information from across the internet to provide direct answers to questions (Google, 2012), and the recent rise of digital personal assistants (e.g., Alexa and Siri) furthers demand to capture factual information from text, that can be queried in response to user questions (for a recent discussion see: Rajpurkar et al., 2016).[13] These projects have amassed an impressive volume of factual knowledge; typically millions of entities (including people, places, organizations, music album,

[12] When extracting facts to populate a knowledgebase from a large source (e.g., the internet or Wikipedia), additional steps may be required to disambiguate between people who share the same name (Cucerzan, 2007). In this dissertation, disambiguation does not pose a substantial issue; since the name of the firm from which a background was obtained is known, and at least within a particular firm, manager names tend to be unique, there is little ambiguity over who a statement refers to.

[13] It should be noted that the word 'fact' is intended here in a relatively narrow sense; most approaches to extract textual information from text tend to be restructured to capturing properties of an entity (e.g., the data firm is founded), or relationships between entities (e.g., the CEO of A is B) (e.g., Suchanek, Kasneci, and Weikum, 2007). Although it is possible to construe any statement as a fact (e.g., for any given sentence in a firm's annual report it is a 'fact' that the firm made that statement, even if the sentence itself is subjective), this is not the intended meaning here.

movies, etc.), each populated with associated properties such as dates, family relations, sizes (e.g., the heights of people) and locations.

However, while 'facts' may form the basis for answering many forms of questions (and especially questions commonly searched for on the internet), the uses of textual data in the social sciences is qualitatively different from merely extracting facts from text. Specifically, there is a much greater interest in analyzing the manner in which information is conveyed (e.g., Bettman and Weitz, 1983; Boje, 1991; Harmon, 2018), as well as using what is said to infer and analyze other concepts, such as how categories form (Kennedy, 2005; Navis and Glynn, 2010), how concepts evolve (Ocasio and Joseph, 2005; Hsu and Grodal, 2015), how identity is presented (Chreim, 2005), and attention patterns (D'Aveni and MacMillan, 1990). These areas of research consider not only what is said, but how it is said, and how what is said changes, with research often drawing from multiple sources to consider how different audiences discuss the same information. Indeed, much social science research involves characterizing the source material, rather than treating the source material merely as a conduit of desired facts.

Despite the breadth of information extraction projects capturing factual properties (e.g., Nakashole, Theobald, and Weikum, 2011; Biega, Kuzey, and Suchanek, 2013), there is little evidence of comparable effort to characterize the source material or to capture subjective information and the structuring of texts. This is likely in part because there is limited demand for characteristics of the source material or subjective information outside of academia; knowing how a manager's experiences are characterized across historical filings has limited applications outside of scholarly research. The predominant focus on factual information may also in part be because while facts have a 'right' answer (e.g., the location of a company's headquarters, that can be returned in response to a question), subjective information does not. Paradoxically, part of

the reason why subjective information can be interesting to social science research, including that it can be framed to present a desired state (e.g., Fiss and Zajac, 2006), and de-coupled from reality (e.g., Meyer and Rowan, 1977), may limit the overall interest of data scientists to extract such relationships.

Overview of the Approach Developed in this Dissertation

This dissertation draws from information extraction research, extending the approaches to allow entire texts to be systematically represented, with particular attention to capture subjective dimensions while preserving the structure of the text.[14] While the developed qualitative and computational approach is general, it is illustrated in the context of managerial backgrounds, allowing each of the stages to be described in a domain with clear research opportunities.[15] The key premise of the approach is that, within a particular communication medium, there is typically substantial deep-level similarity in the dimensions of discussion; there is a large degree of similarity in broad areas discussed in managerial backgrounds. Overlayed, however, is

[14] While this dissertation draws on information extraction approaches (e.g., Suchanek, Kasneci, and Weikum, 2007; Gupta and Lehal, 2009; Han, Pei, and Kamber, 2012: 113) and recent advancements in machine learning (e.g., Hochreiter and Schmidhuber, 1997; Srivastava et al., 2014; Lample et al., 2016) since systematically capturing a representation of the text is qualitatively different from capturing specific facts, the approach developed also adapts and extends the approaches. Specifically, particular qualitative consideration is given to the underlying material, considering the dimensions on which sentences vary, and how to represent entire sentences in ontologies to represent the entire meaning of what is said. The approach also extends the concepts that are captured to the relative domain-specific and subjective information characteristic of managerial backgrounds (e.g., management titles, committees, company descriptions, etc.), developing standardization approaches to facilitate comparisons and connections to external databases. The need to capture relatively domain-specific concepts also means that, in commonality with other domain-specific information extraction approaches (see, e.g., Cohen and Hersh, 2005), the ability to draw on existing 'general purpose' information extraction approaches is limited. Thus, while there are many tools for extracting certain common types of information from text, for example people names, company names, dates, locations, contact information such as email addresses and phone numbers, and sometimes products ((e.g., OpenNLP: Baldridge, 2005; NLTK: Bird and Loper, 2004; and Stanford NER: Finkel, Grenager, and Manning, 2005), in order to capture the specific information desired, the actual implementation is custom; all of the code is custom written (with the exception of various OpenSource standard Python libraries such as Google TensorFlow: Abadi et al., 2016).

[15] While the data source used in this dissertation comprises approximately 8 million sentences taken from the proxy statements of US public firms across the period 2007-2017, the ontologies and population approach are intended to serve as the basis for extension across other communications mediums, such as company websites, and extended time periods.

substantial surface-level variation, with names, acronyms, synonyms, and slight differences in sentence construction introducing substantial variation into the sentences, while making little to no difference to the underlying meaning. These surface level variations mean that in the context of top manager experiences, the only occurrences of identical sentences are year-to-year re-use of sentences for the same manager.[16] By identifying the underlying similarities on which the sentences are comparable, ontologies can be developed to represent the information, while abstracting surface-level variations.

Through careful qualitative consideration of managerial backgrounds, five primary information types were identified: background information, typically not the focus of a sentence, such as a manager's name or age (`BACKGROUND`), a manager's current/past positions (`POSITIONS`), experiences that a manager has gained (`EXPERIENCES`), qualifications that a manager has received (`QUALIFICATIONS`), and professional licenses that a manager holds (`PROFESSIONAL_LICENSES`).[17] Each of these primary information types captures qualitatively different types of discussion, and together represent substantively all of the information in managerial backgrounds. As will be discussed, by systematically considering the dimensions on which sentences in these primary information types are comparable, it is possible to successively parse each information type, progressively forming the basis of the ontologies. Through this process, the standardized ontologies can be developed that preserve the meaning of what is said, and the structuring of that meaning, while abstracting surface-level variations, such that it is possible to directly compare the texts on dimensions of interest. Using information extraction techniques, that are extended to characterize entire sentences including subjective information, this dissertation illustrates how

[16] This is was true in various other domains explored as part of the dissertation process: essentially every sentence in on technology company websites; description of business; and risk statements were unique, with the exception of occasional boilerplate (e.g., 'Contact us for more information').

[17] Full details of the design and specification of the ontologies, development approach to populate the information, and envisioned theoretical contributions of the approach are given in due course.

the developed ontologies shown in Figures I.2-4 can be systematically populated, allowing a consistent, standardized, representation of what is said to be captured across the population of managers.[18]


```
{
  "ORIGINAL_SENTENCE":"From October 2001 to November 2004, she served as Vice President of Operations and a
                       director for QRS Corp., a gold mining company, and between March 1996 and May 2001 was the
                       CEO of Vaynol Clothing, a leading US retailer of women's clothing",
  "POSITIONS":[
   { "ORIGINAL":"From October 2001 to November 2004, she served as Vice President of Operations and a director of
                 QRS Corp., a retail supply chain software and services company",
     "START_DATE":{"ORIGINAL":"October 2001","YEAR":2001,"MONTH":10},
     "END_DATE":{"ORIGINAL":"November 2004","YEAR":2004,"MONTH":11},
     "JOB_TITLES":[{"ORIGINAL":"Vice President of Operations","LEVEL":"VICE_PRESIDENT","AREA":["OPERATIONS"]},
                   {"ORIGINAL":"Director","LEVEL":"DIRECTOR"}],
     "COMPANY":{"ORIGINAL":"QRS Corp.","CLEANED":"QRS"},
     "COMPANY_DESCRIPTION":{"ORIGINAL":"NYSE-listed gold mining company",
                            "LISTING-OWNERSHIP":{"OWNERSHIP_TYPE":"PUBLICALLY_LISTED",
                                                 "EXCHANGE":[{"EXCHANGE_NAME":"NYSE","COUNTRY":"USA"}]},
                            "INDUSTRY":{"NAICS_3_DIGIT":212
                                        "NAICS_3_DESCRIPTION":"Mining (except oil and gas)"}},
   { "ORIGINAL":"between March 1996 and May 2001 was the CEO of Vaynol Clothing, a leading US retailer of women's
                 clothing",
     "START_DATE":{"ORIGINAL":"March 1996","YEAR":1996,"MONTH":3},
     "END_DATE":{"ORIGINAL":"May 2001","YEAR":2001,"MONTH":5},
     "JOB_TITLES":[{"ORIGINAL":"CEO","LEVEL":"CEO"}],
     "COMPANY":{"ORIGINAL":"VAYNOL Clothing","CLEANED":"VAYNOL COTHING"},
     "COMPANY_DESCRIPTION":{"ORIGINAL":"leading US retailer of women's clothing",
                            "INDUSTRY":{"NAICS_3_DIGIT":448,
                                  "NAICS_3_DESCRIPTION":"Clothing and Clothing Accessories Stores"},
                            "REGION":[{"COUNTRY":"USA"}],
                            "CHARACTERIZATION":[{"TERM":"leading","AREA":"LEADING"}]}
}
```


Figure I.2. Example of a Sentence Discussing a Manager's Position History Represented in a Standardized Manner[19]

[18] The hierarchical (JSON) structure shown here is increasingly used to transfer information between organizations, and is very flexible, allowing a hierarchical structure to be defined as needed.

[19] All examples of managerial backgrounds contained in this dissertation are pseudo-examples, written to convey the form and the style of the underlying material, while not directly based on any one example.

```
{
  "ORIGINAL_SENTENCE":"Mr. Doe earned a Bachelor degree in engineering from the University of
                       Michigan and a Master of Business Administration from the Stanford Graduate School of
                       Business.",
  "BACKGROUND":{"PERSON_NAME":{"ORIGINAL":"Mr. Doe","NAME_TITLE":"MR","LAST_NAME":"DOE"}
  "QUALIFICATIONS":[
      { "EDUCATION_INSTITUTION":{"ORIGINAL":"University of Michigan","UNIVERSITY":"UNIVERSITY OF MICHIGAN"},
        "DEGREE":{"ORIGINAL":"Bachelor degree in engineering",
                  "LEVEL":"UNDERGRADUATE/BACHELORS","SUBJECTS":["ENGINEERING"]}},

      { "EDUCATION_INSTITUTION":{"ORIGINAL":"Stanford Graduate School of Business"
                       "UNIVERSITY":"STANFORD UNIVERSITY","DEPARTMENT":"Graduate School of Business"}
        "DEGREE":{"ORIGINAL":"Master of Business Administration"
                  "LEVEL":"GRADUATE/MASTERS","SUBJECTS":["BUSINESS"]}}
   ]
}
```


Figure I.3, Example of a Sentence Discussing a Manager's Qualifications Represented in a Standardized Manner


```
{
  "ORIGINAL_SENTENCE":"Jone Doe has spent 5 years working in the automotive industry",
  "BACKGROUND":{"PERSON_NAME":{"ORIGINAL":"Mr. Doe","NAME_TITLE":"MR","LAST_NAME":"DOE"}
  "EXPERIENCE":[{"ORIGINAL":"worked in the automotive industry for 5 years"
                 "LENGTH_OF_TIME":{"ORIGINAL":"5 years","UNIT":"YEAR","QUANTITY":5}
                 "AREAS":{"ORIGINAL":"automotive industry","NAICS_3_DIGIT":336
                          "NAICS_3_DESCRIPTION":"Transportation Equipment Manufacturing"}
                 "EXPERIENCE_TYPE":{"ORIGINAL":"worked","TYPE":"WORK"}
}
```


Figure I.4, Example of a Sentence Discussing a Manager's Experiences Represented in a Standardized Manner


```
{
  "ORIGINAL_SENTENCE":"Mr. Doe is licensed to practice law in the State of Michigan",
  "BACKGROUND":{"PERSON_NAME":{"ORIGINAL":"Mr. Doe","NAME_TITLE":"MR","LAST_NAME":"DOE"}
  "PROFESSIONAL_LICENSE":[
      {
         "ORIGINAL":"licensed to practice law in the State of Michigan",
         "AREA":"LAW",
         "REGIONS":[{"COUNTRY":"USA","STATE":"MI"}]
      }]
}
```


Figure I.5. Example of a Sentence Discussing a Manager's Professional License Represented in a Standardized Manner

As well as describing an approach to develop and populate the ontologies, various ways of assuring the validity of the populated ontologies are introduced and discussed. These include: i) significant qualitative consideration to develop the ontologies; ii) validation of properties by dissecting terms to their components; iii) validation of terms through external data-checks (such

as university names, and locations); iv) validation of concepts by the context in which they appear in the text; v) manual oversight to ensure the ontologies are populated as expected; and vi) illustration of terms in each concept, to provide face validity of the classifications.[20] Other fully and semi-automated approaches to help ensure terms are appropriately categorized, and categorizations of terms validated, are discussed, including an automated approach to connect industries to associated NAICS-codes irrespective of whether the term directly appears in the NAICS classification manual, using surrounding discussion of the terms on Wikipedia. The series of validation approaches are intended to increase confidence in the populated ontologies, and have applications beyond this dissertation, broadly facilitating theoretically-centered research inherently requiring large volumes of rich, nuanced textual data, where it is unfeasible to manually verify every term individually.

The dissertation also considers specific research questions facilitated by systematically representing information in standardized ontologies. As will be discussed, there are several layers of opportunity, the first arising from the ability to precisely measure very specific relationships, such as how the descriptions of the firms that a manager has worked for are adjusted between years. By first representing the information in a consistent manner, it becomes feasible to capture specific changes of interest, including differences that may be hard to manually code, even on a relatively small scale. For example, the ontologies allow blocks of text to be 'subtracted' from one another, allowing consideration of how particular sentences, in much larger documents, change from year to year (even if the order of the sentences in the documents differ). Being able to capture such specific differences also helps enable comparisons between mediums, despite the presence of extraneous variation unrelated to the dimension of interest.

[20] Going forward, the face-validity is intended to be built into the approach, with classification summaries providing high overall transparency to the approach.

Being able to systematically and consistently represent textual information in ontologies also facilitates research considering multiple layers of information structuring. By preserving the order of information within sentences, and the overall order of sentences, it is possible to characterize the multiple layers of structuring, including how sub-structures combine to overall meta-structures. Managerial backgrounds are well suited for considering the layers of structuring, allowing separate consideration of how information is structured within a managerial background, and how the backgrounds of individual managers combine at the top management team level. The ontologies make direct examination of the evolution of the different levels of structuring feasible, including consideration of subtle changes that may be hard to manually code, such as adjustments over time to the ordering of information.

Being able to systematically characterize entire populations of text also allows greater identification and awareness of the occurrence of gradual societal changes. A large proportion of organizational theories are 'mid-range' in nature (Merton, 1957), grounded in the real world phenomenon that they help explain (e.g., research on poison pills: Davis, 1991; acquisitions: Haunschild, 1993; adoption of TQM practices: Westphal, Gulati, and Shortell, 1997; or the difficulties faced by females and racial minorities in organizations: Eagly and Carli, 2007). Although theory may guide our understanding of these areas, theoretical development rarely occurs in a vacuum, and studying and explaining organizational change requires at least an awareness of the occurrence of the phenomenon itself (Eisenhardt and Graebner, 2007). While certain organizational changes may be readily apparent to researchers due to the suddenness of the change (e.g., the rapid diffusion of poison pills) or societal-level discussion of the issue (e.g., discrimination in the work-place), much societal change occurs gradually, garnering little main-stream attention. This may be particularly true for changes to corporate management practices,

where the sheer volume of textual information in corporate filings, external communications, and conference calls, make it inherently difficult to for any individual to synthesize. Indeed, the 'black-box' of corporate governance (e.g., Daily, Dalton, and Cannella, 2003) may be as much a factor of the excessive volume of available information, rather than an absence. Systematically standardizing textual information allows gradual and nuanced changes in corporate leadership to be more easily identified, providing new contexts for theoretical development.

Overview of the Dissertation

The remainder of this dissertation develops and describes the approach to capture and represent meaning from textual information, and the theoretical opportunities enabled by doing so, with a particular focus on managerial backgrounds. Chapter 2 illustrates the largely qualitative approach by which the ontologies were developed. Chapter 3 describes each of the three key stages in the information extraction process: i) concept identification, ii) interpretation of meaning from semantic relationships, and iii) information standardization. This is then extended in Chapter 4, which explains in greater detail how key challenges in systematically transforming the managerial backgrounds to the developed ontologies were addressed. Chapter 5 describes aggregation and comparison approaches, with consideration of theoretical dimensions that the ontologies facilitate examining. This discussion continues in Chapter 6, with greater consideration of specific questions enabled in the context of managerial backgrounds. Finally, the dissertation concludes in Chapter 7 with a discussion of overall contributions, illustrating further theoretical opportunities from extending the approach to capture a broader array of meaning.

CHAPTER II

Qualitative Development Process

This chapter explains the qualitative approach by which the ontologies in this dissertation were developed. The approach draws from research on identifying themes, considering similarity and differences to identify relevant dimensions on which texts vary (Glaser and Strauss, 1967; Ryan and Bernard, 2003). As described further below, this process began by reading a substantial number of managerial backgrounds from the proxy statements of US public firms from 2007-2017, to understand the nature of the material, and continued until theoretical saturation was reached, where further reading gave little additional insight (Strauss and Corbin, 1990). This provided the basis to initially classify the sentences into five primary information types (`BACKGROUND, POSITIONS, EXPERIENCES, QUALIFICATIONS, PROFESSIONAL_LICENSES`), and continued until ontologies for each information type were developed. These initial ontologies were then extended during the implementation process to allow less common dimensions to also be captured. Over this process, many thousand managerial backgrounds were read in whole or in part, providing a high level of familiarity with the material.

Unit of Analysis and Primary Information Types

The first consideration in representing the information was deciding an appropriate unit of text to represent. While sentences in some communication mediums use pronouns to refer to previously introduced terms, the sentences in managerial backgrounds were almost entirely self-

contained, with very few references to terms in prior sentences.[21] Since each sentence could be interpreted independently, the sentence was taken as the unit of analysis (Weber, 1985: 22). Thus, while Chapter 5 will discuss how it is possible to aggregate individual sentences at the manager and top management team level, the remainder of this and the next two chapters will focus on capturing representations of individual sentences.

Just as résumés tend to cover similar material (McGrimmon, 2014), there is inherent similarity in the material discussed in managerial backgrounds. By considering the themes underlying each sentence (Glaser and Strauss, 1967; Ryan and Bernard, 2003), five commonly occurring primary information types were relatively easily identified. The first information type, **BACKGROUND**, included details such as name, age, and salutation, that while typically not the focus of the sentence, indicated who the description referred to. This information was usually included in the first sentence of a manager's background, with subsequent sentences referring to the manager as '**he**', '**she**' or by first name. The second information type, **POSITIONS**, described a manager's current and prior positions, including board and committee appointments. This description typically included job titles, dates of employment, names of the firms worked for, and often a short description of those firms (e.g., '… **he served as vice president of Gorilla Parts, *a leading NYSE-listed manufacturer of automotive components***'). The third information type, **EXPERIENCES**, discussed the more subjective elements of a manager's experiences, including the industries and functional areas worked in, and the length of these experiences. The fourth information type, **QUALIFICATIONS**, described the degrees that a manager had obtained, including graduation dates, the name of the granting universities, and usually the

[21] The only substantial use of pronouns was the use of personal pronouns to refer to a manager after the first sentences (e.g., '**he**', '**she**'), and use of the terms such as '**the company**' or '**our**' (e.g., '**has worked for the company since 1996**', or '**has served on our board since 1996**'). Neither of these pose a problem to interpreting the meaning; the subject of personal pronouns is always the focal manager, and the company is likewise the focal company, both of which are known for sentences under consideration.

subjects of the degrees. The final information type, **PROFESSIONAL_LICENSES**, discussed managers' professional licenses, often including the issuing state or country (e.g., '**Certified Professional Accountant**', or '**registered engineer in the State of Michigan**').

As illustrated in Table II.1, with the exception of background details (**BACKGROUND**), each information type was typically the focus of an entire sentence, and each discussed relatively distinct material. While there are some similarities between **QUALIFICATIONS** and **PROFESSIONAL_LICENSES** they are usually discussed in separate sentences and have quite different dimensions; professional licenses were often associated with specific states, may expire, and while not noted in the text, often have legal implications.

BACKGROUND	Mr. John Doe (age 56) is the…
	Mrs. Doe is the…
	Dr. Doe, Ph.D., is the…
	Mr. John Doe III has worked…
POSITIONS	Mr. Doe served as Executive Vice President and Chief Financial Officer at Gorilla Software from September 1997 until his retirement in January 2003.
	From 1992 to 1993, Mr.Doe was Vice President of Business Development at Gorilla Software, a clinical research organization in Detroit, MI.
	He has been a Director of the Company since 2002 and chairman of the Compensation Committee since 2005.
	Prior to joining Oculi Machined Parts, Ms. Doe was the Executive Vice President for Gorilla Software, a company developing speech recognition software.
	Mr. Doe has been Vice President, Secretary, Treasurer and a director of Gorilla Software since 1980 and held other positions with Oculi Machined Parts prior thereto.
EXPERIENCES	Mr. Doe also brings noteworthy business sector executive officer experience.
	Mr. Doe also has a solid understanding of the Company's electric operations and the Florida market.
	He has over 35 years of experience in the information technology and security marketplace.
	Mr. Doe has over 20 years of investment industry experience.
	Mr. Doe has over 35 years of experience in finance in the banking industry.
QUALIFICATIONS	He also holds a BBA from the Ross School of Business at the University of Michigan and holds a Juris Doctorate from Yale University.
	Dr. Doe received his Ph.D. in Biochemistry from Case Western Reserve University in Cleveland, Ohio and his B.A. with honors in Chemistry from Taylor University in Upland, Indiana.

	Mr. Doe has an undergraduate degree in accounting from DePaul University and an MBA from the University of Chicago.
	He graduated from Massachusetts Institute of Technology with a Bachelor's degree in Mechanical Engineering, and he earned an MBA in General Management from Harvard Graduate School of Business in 1968.
	He is a graduate of California State University, Northridge, with a B.S. degree in accounting.
	Mr. Doe received a Bachelor's degree and a Master's degree in Accounting from Tianjin University of Finance and Economics.
PROFESSIONAL_LICENSES	Ms. Doe is a Certified Public Accountant.
	Mr. Doe is a licensed attorney in Colorado, New York, and Texas.
	He is a Certified Public Accountant and is licensed to practice law in the Commonwealth of Puerto Rico.
	He is a California licensed certified public accountant.

Table II.1 Illustration of Sentences Typical of Each Information Type.

Development of the Primary Ontologies

The five information types form the basis of the overall standardized representation. As illustrated in Figure II.1, it is possible to cleanly split sentences into these information types.

```
a)  {
   "ORIGINAL_SENTENCE":"John Doe was the CFO of Gorilla Software from 1997 until 2000 and then was the Vice
                        President of Finance for Bridge Analysis from 2001 until March 2008",
   "BACKGROUND": -- John Doe --
   "POSITIONS": -- CFO of Gorilla Software from 1997 until 2000 --
               -- Vice President of Finance for Bridge Analysis from 2001 until March 2008 --
}

b)  {
   "ORIGINAL_SENTENCE":"John Doe received a BA in Mathematics from the University of Michigan in 1996, an
                        MBA from Stanford in 2001, and most recently became licensed as a certified public
                        accountant in the state of California in 2007,
   "BACKGROUND": -- John Doe --
   "QUALIFICATIONS": -- BA in Mathematics from the University of Michigan in 1996 --
                     -- MBA from Stanford in 2001 --
   "PROFESSIONAL_LICENSES": -- became licensed as a certified public accountant in the state of California
                               in 2007 --
}

c)  {
   "ORIGINAL_SENTENCE":"John Doe has substantial financial experience, with a particular emphasis on valuing
                        mergers and acquisitions",
   "BACKGROUND": -- John Doe --
   "EXPERIENCES": -- substantial financial experience, with a particular emphasis on valuing mergers and
                     acquisitions --
}
```

Figure II.1. Illustration of Sentences Split into the Five Primary Ontologies

The examples in Figure II.1 also illustrate several important characteristics of the sentences. First, sentences can contain more than one type of information, commonly just BACKGROUND and one other, although occasionally more. For example, the sentence a) includes information in the BACKGROUND and the POSITIONS types, b) includes information in the BACKGROUND, QUALIFICATIONS, and PROFESSIONAL_LICENSES types, and c) includes information in the BACKGROUND and EXPERIENCES types. Second, as illustrated in examples a) and b), there may be more than one distinct part within each information type; there are two distinct positions mentioned in example a) and two distinct qualifications mentioned in example b). Both of these characteristics are reflected in the specification of the overall sentence representation shown in Figure II.2. This representation provides the basis for dissecting the sentences and shows how, with the exception of BACKGROUND, each of the primary-ontologies comprises a list of sub-ontologies, representing distinct positions, qualifications, experiences and professional licenses, in the order that they appear in the sentence.[22]

```
{
   "ORIGINAL_SENTENCE": Original text
   "BACKGROUND": All background details discussed (e.g., person name, age)
   "POSITIONS": List of POSITION sub-ontology
   "EXPERIENCES": List of EXPERIENCE sub-ontology
   "QUALIFICATIONS": List of QUALIFICATION sub-ontology
   "PROFESSIONAL_LICENSES": List of PROFESSIONAL_LICENSE sub-ontology
}
```

Figure II.2. Overall Sentence Representations

There was also substantial commonality in the dimensions discussed within each information type. The approach used to represent the meaning within the sub-ontologies was thus to identify the commonly occurring concepts, dissecting each component separately. For

[22] Since backgrounds focus on a single manager, there is no need to structure the BACKGROUND primary-ontology as a list. While the representation allows all five primary ontologies to be captured from a particular sentence, there are no instances where all five information types are simultaneously discussed in a single sentence. Similarly, while the structure allows an unlimited number of separate positions, experiences, qualifications and professional licenses to be captured, typically this will be a small number.

example, in the text describing an individual **POSITION**, such as '**Chief Financial Officer of Gorilla Software from 1997 until 2000**', the job title ('**Chief Financial Officer**'), company name ('**Gorilla Software**'), start date ('**1997**') and an end date ('**2000**') can easily be recognized.[23] Similarly, in the text describing an individual **QUALIFICATION**, such as '**MBA from the University of Michigan in 1996**', the degree ('**MBA**'), the university ('**University of Michigan**') and the graduation date ('**1996**') can again be readily identified. The underlying similarity within information types made it possible to develop the **POSITION**, **QUALIFICATION**, **EXPERIENCE**, and **PROFESSIONAL_LICENSE** sub-ontologies in Figure II.3.

```
POSITION:{
   "ORIGINAL":"Chief Financial Officer of Gorilla Software Inc. from April 1997 until March 2000",
   "JOB_TITLE": -- Chief Financial Officer --
   "COMPANY_NAME" -- Gorilla Software --
   "START_DATE": -- April 1997 --
   "END_DATE": -- March 2000 --
}

EXPERIENCE:{
   "ORIGINAL":"substantial financial experience, with a particular emphasis on valuing mergers and
              acquisitions",
   "CHARACTERIZATION": -- substantial --
   "AREAS": -- financial experience, with a particular emphasis on valuing mergers and
              acquisitions --
}

QUALIFICATION:{
   "ORIGINAL":"BA in Mathematics from the University of Michigan in 1996",
   "QUALIFICATION": -- BA in Mathematics --
   "EDUCATION_INSTITUTION": -- University of Michigan --
   "GRADUATION_DATE": -- 1996 --
}

PROFESSIONAL_LICENSE:{
   "ORIGINAL":"licenced CPA in the state of Michigan",
   "LICENSE": -- CPA --
   "LOCATION": -- state of Michigan --
}
```

Figure II.3. Managerial Position in the **POSITION** Sub-Ontology

[23] Certain concepts such as job titles, company names, dates correspond to natural concepts, or concepts that are widely shared and easily recognized (Rosch, 1973; Murphy, 2002). While other concepts are slightly more abstract such as '**experiences**', the engagement with the material again indicated that this was a relatively meaningful basis on which to characterize the text.

The process of simplifying continued until all sub-components were dissected to the most basic element. The dimensions of variation were again identified by considering similarities and differences across a large number of items of a particular type (e.g., Glaser and Strauss, 1967). For example, as indicated in Figure II.4, functional area and seniority level were identified as two meaningful dimensions on which managerial backgrounds could be characterized; two dimensions that as well as appearing qualitatively distinct, are also considered separately within the academic literature (e.g., functional area: Michel and Hambrick, 1992; Ocasio and Kim, 1999; and seniority level: Wooldridge and Floyd, 1990; Radner, 1992).

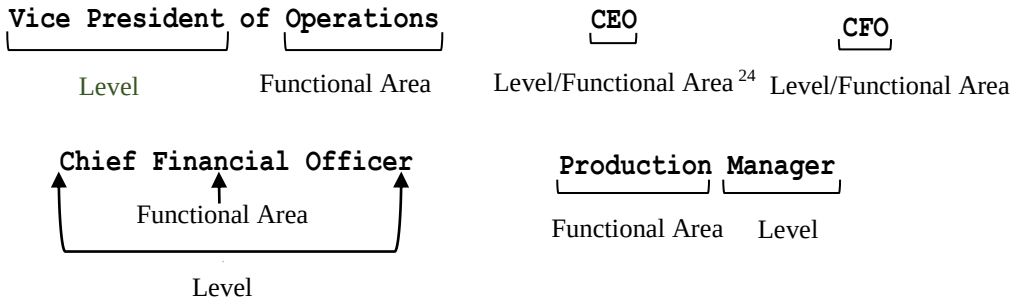


Figure II.4. Illustration of Managerial Titles Dissected by Key Dimensions[24]

A small number of dimensions characterized information types: qualifications can typically be characterized by level and subject; dates by day, month, year; and people names by title, first name, last name, and salutation. Identifying the key dimensions also facilitated identifying less common, although qualitatively distinct, dimensions of variation. For example, managerial titles sometimes included a region, either specifically (e.g., '**Head of Operations in China**') or more generally (e.g., '**Head of International Sales**'). Since the region that a manager works in is qualitatively different from their functional area (e.g., Allred, Snow, and

[24] There are a limited number of positions, including '**CEO**' and '**Director**', where the functional area is more general and integral to the layer (i.e., with the overall management or oversight of the firm); the functional area for such area for such positions are currently classed as '**GENERAL/OVERSIGHT**'.

Miles, 1996) it was considered a distinct dimension. Similarly, the managerial titles sometimes included the industry in which the manager worked ('**Head of Pharmaceutical Operations**'), which is also a qualitatively distinct dimension from functional area, level, and region. Moreover, during the implementation stage (described in Chapter 3 and 4) it was possible to identify much less common dimensions of the text. For example, by reviewing the parts of the management title not captured within the initial sub-ontology across the population of identified managerial titles (e.g., from 2007-2017) other dimensions discussed in the literature, including '**interim**' and '**co-**', were identified (Ballinger and Marcel, 2010; Krause, Priem, and Love, 2015). Full details of all sub-ontologies, including possible values that properties can take, discussion of how unusual cases are handled, and examples of populated ontologies are included in Appendix D. As described later, the ontologies also ensure that terms such as '**CEO**' and '**Chief Executive Officer**', and '**BBA**' and '**Bachelors in Business Administration**' are treated the same. Figure II.5 shows examples of the **POSITION**, **EXPERIENCE**, **QUALIFICATION**, **PROFESSIONAL_LICENSE** sub-ontologies.


```
POSITION:{
   "ORIGINAL":"Chief Financial Officer of Gorilla Software Inc. from April 1997 until March 2000",
   "JOB_TITLE":{"ORIGINAL":"Chief Financial Officer","LEVEL":"CHIEF-OFFICER","AREA":["FINANCE"]}
   "COMPANY_NAME":{"ORIGINAL":"Gorilla Software","NAME_CLEANED":"GORILLA SOFTWARE"}
   "START_DATE":{"ORIGINAL":"April 1997","YEAR":1997,"MONTH":4}
   "END_DATE":{"ORIGINAL":"March 2000","YEAR":2000,"MONTH":3}
}

EXPERIENCE:{
   "ORIGINAL":"worked in the automotive industry for 5 years"
   "LENGTH_OF_TIME":{"ORIGINAL":"5 years","UNIT":"YEAR","QUANTITY":5}
   "AREAS":{"ORIGINAL":"automotive industry","NAICS_3_DIGIT":336
            "NAICS_3_DESCRIPTION":"Transportation Equipment Manufacturing"}
   "EXPERIENCE_TYPE":{"ORIGINAL":"worked","TYPE":"WORK"}
}

QUALIFICATION:{
   "ORIGINAL":"received a BEng in 1990 with a from the University of Illinois"
   "DEGREE":{"ORIGINAL":"BEng","DEGREE_LEVEL":"UNDERGRADUATE/BACHELORS","SUBJECTS":["ENGINEERING"]}
   "EDUCATION_INSTITUTION":{"ORIGINAL":"University of Illinois","UNIVERSITY":"UNIVERSITY OF ILLINOIS"}
   "GRADUATION_DATE":{"ORIGINAL":"1990","YEAR": 1990}
}
```

```
PROFESSIONAL_LICENSE:{
   "ORIGINAL":"licensed to practice law in New Jersey",
   "AREA":"LAW",
   "REGION_LIST":[{"COUNTRY":"USA","STATE":"NJ"]
}
```


Figure II.5. Managerial Position Standardized in the **POSITION**, **EXPERIENCE**, **QUALIFICATION**, and **PROFESSIONAL_LICENSE** Sub-Ontologies

Removal of Surface Level Variation

Once the text has been standardized, the four sources of surface-variation shown below in Table II.2 are either removed or captured. Specifically permutations in sentence constructions, acronyms and synonyms are removed, and label names (e.g., people names) are captured, such that comparisons can be made on dimensions of interest, ignoring label names as appropriate.[25]

Surface-variation type	Example
Acronyms	The *CEO* of Gorilla Software is John Doe *vs.* The *Chief Executive Officer* of Gorilla Software is John Doe
Synonyms	John Doe *works for* Gorilla Software *vs.* John Doe *is employed by* Gorilla Software
Slight variations in sentence construction[26]	The CEO of Gorilla Software is John Doe *vs.* John Doe is the CEO of Gorilla Software
Differences in specific labels (e.g., people and organization names)	*John Doe* is the CEO of *Gorilla Software* *vs.* *Jane Roe* is the CEO of *Oculi Machined Parts*

Table II.2. Summary of Surface-Level Variation

[25] As explained in greater detail later, while the sentence seeks to represent all dimensions of what is said, it is likely that only one or two parts will be of interest to a particular research question; any dimensions not relevant may be ignored. This is similar to how, while COMPUSTAT has thousands of financial variables about organizations, allowing the spectrum of questions concerning financial information to be examined, typically only a handful of variables are of relevance to a particular research question.

[26] It should be noted that while effort has been made to ensure that key dimensions of ordering within the sentence are preserved (e.g., the order experiences are introduced, the order positions are introduced, etc.), the process of standardizing the sentence inherently, and intentionally, abstracts some slight within-sentence variations in order. For example, '`from 1997 to 2000 he was CEO`' and '`he was CEO from 1997 and 2000`' standardize the same. While it is hard to see how, at least for the vast majority of possible research question, that this could be a meaningful difference, it would be is feasible at some point in the future to have an option to number the position in the sentence of every component extracted, allowing very slight within-sentence variations to be captured, should they be desired. To date though, this has not been added since it makes the structures much harder to understand and explain, while offering very little value.

To illustrate the power of this process to abstract surface-level variations, Figure II.6 provides sentences that while superficially varying from those in Figure II.5 (i.e., acronyms, synonyms, different label names and slight differences in sentence constructions), standardize the same on all properties except the original text, and label names.


```
POSITION:{
   "ORIGINAL":"from April 1997 through March 2000 was the Chief Finance Officer of Bridge Analysis",
   "JOB_TITLE":{"ORIGINAL":"Chief Finance Officer","LEVEL":"CHIEF-OFFICER","AREA":["FINANCE"]}
   "COMPANY_NAME":{"ORIGINAL":"Bridge Analysis","NAME_CLEANED":"BRIDGE ANALYSIS"}
   "START_DATE":{"ORIGINAL":"April 1997","YEAR":1997,"MONTH":4}
   "END_DATE":{"ORIGINAL":"March 2000","YEAR":2000,"MONTH":3}
}
ALTERNATIVE_ABOVE:"                                                                                ",

EXPERIENCE:{
   "ORIGINAL":"spent 5 years working in the automotive sector"
   "LENGTH_OF_TIME":{"ORIGINAL":"5 years","UNIT":"YEAR","QUANTITY":5}
   "AREAS":{"ORIGINAL":"automotive sector","NAICS_3_DIGIT":336
            "NAICS_3_DESCRIPTION":"Transportation Equipment Manufacturing"}
   "EXPERIENCE_TYPE":{"ORIGINAL":"working","TYPE":"WORK"}
}
ALTERNATIVE_ABOVE:"worked in the automotive industry for 5 years"

QUALIFICATION:{
   "ORIGINAL":"earned an undergraduate degree in engineering from the University of Illinois in 1990"
   "DEGREE":{"ORIGINAL":"BEng","DEGREE_LEVEL":"UNDERGRADUATE/BACHELORS","SUBJECTS":["ENGINEERING"]}
   "EDUCATION_INSTITUTION":{"ORIGINAL":"University of Illinois","UNIVERSITY":"UNIVERSITY OF ILLINOIS"}
   "GRADUATION_DATE":{"ORIGINAL":"1990","YEAR":1990}
}
ALTERNATIVE_ABOVE:"                                                                  ",

PROFESSIONAL_LICENSE:{
   "ORIGINAL":"licensed in the State of New Jersey to practice law",
   "AREA":"LAW",
   "REGION_LIST":[{"COUNTRY":"USA","STATE":"NJ"]
}
ALTERNATIVE_ABOVE:"                                        ",
```


Figure II.6. Illustration of How Different Sentence Constructions Standardize the Same

CHAPTER III

Overview of the Information Extraction Process

This chapter introduces the three stages of populating the ontologies: i) concept recognition, ii) semantic interpretation, and iii) standardization, using simple, stylized examples to explain each stage. Chapter 4 then builds on these examples to explain how the approach was expanded to the complex and varied sentence structures common in managerial backgrounds, with greater consideration to validate the information.

Stage 1: Concept Recognition

The process by which the information is populated into the ontologies has direct parallels with the process by which the ontologies were developed. While the ontologies were developed by systematically dissecting the sentence to underlying concepts, the approach to populate these ontologies begins by mapping words in the sentence to underlying concepts.[27] Moreover, just as the similarity in the dimensions allowed the ontologies to be developed, reducing the text to concepts substantially simplifies the sentences. For example, despite sharing little common words, the two sentences '`John Doe is a CEO of Gorilla Software`', and '`Jane Roe is the Vice President of Oculi Parts`' both reduce to the underlying concepts `PERSON_NAME IS MANAGEMENT_TITLE OF COMPANY_NAME`.

[27] Greater consideration will be given in Chapter 4 of the level of the granularity of the initial concepts captured; for the time being the term 'concept' corresponds to what could be considered 'natural concepts' (e.g., Rosch, 1973; Murphy, 2002) such as `MANAGEMENT_TITLE, DATES, PERSON_NAME, COMPANY_NAME, DEGREE, EDUCATION_INSTITUTION`, and are typically the lowest level of sub-ontologies.

These concepts (e.g., `PERSON_NAME, DATES, COMPANY_NAME`) broadly correspond to natural categories, or concepts that are widely shared and easily recognized (Rosch, 1973; Murphy, 2002), and being able to read and comprehend sentences depends on our ability to identify these underlying concepts (Bower and Trabasso, 1963; Swinney, 1979; Perfetti and Hart, 2001). As such, we can readily identify in the sentence '`John received a BBA from the University of Michigan`' that '`John`' is the name of a `PERSON`, '`BBA`' is a `QUALIFICATION` and '`University of Michigan`' is a university, or `EDUCATION_INSTITUTION`.[28] Our ability to identify patterns in text means that we can also recognize concepts even if we are not familiar with the words. For example, in the sentence '`Shareve received a BBA from Panear Tech`', we can readily deduce that '`Shareve`' is likely a person and that '`Panear Tech`' is likely an `EDUCATION_INSTITUTION`, despite not necessarily being familiar with either. Like people, computers are adept at pattern recognition, and despite lacking a direct understanding of what concepts such as people or universities are, can be trained to recognize the concepts in the text (e.g., Lample et al., 2016).

Stage 2: Semantic Interpretation

While concept identification helps reduce the variation between sentences, the second key stage, semantic interpretation, involves deducing the meaning based on the structure of concepts and the connecting words. For example, in the sentence '`John received a BBA`', by recognizing the `PERSON_NAME` concept and the `QUALIFICATION` concept, connected with the term '`received`', it is possible to deduce that a person called John received a qualification of type BBA. While information extraction is often implemented by interpreting relationships between

[28] While the vast majority of qualifications discussed in managerial backgrounds are from Universities, there are some community colleges; as such a more general term is used throughout.

just two concepts in a sentence (e.g., Nakashole, Theobald, and Weikum, 2011; Biega, Kuzey, and Suchanek, 2013), as illustrated in Figure III.1, it is possible to extend the approach to populate the ontologies by interpreting the sequencing of several concepts.

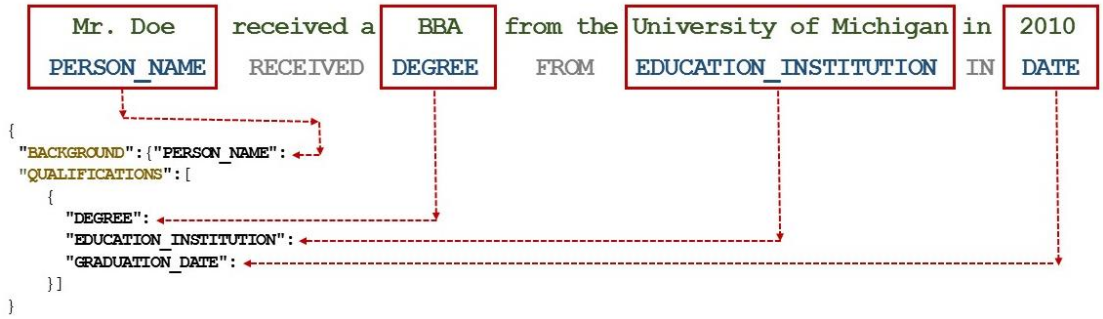


Figure III.1 Illustration of the Process by which Information is Populated to the Ontologies

Thus, although computers do not have a direct understanding of the connecting words, they can use rules to capture the relationships and populate the ontologies. Moreover, other connecting words such as '**obtained**' or '**achieved**' can be interpreted in this context the same as '**received**'. This illustrates why reducing the text to concepts simplifies the semantic interpretation; while it is possible to write semantic rules to populate the ontologies at the concept level, since there are so many possible people names, university names and different ways of saying '**received**', it is unfeasible to construct such rules at the word level. By first reducing the sentence to underlying concepts, the process of interpreting the meaning based on the sequencing of these concepts is greatly simplified.

Stage 3: Standardization

While the first two stages, entity recognition and semantic interpretation, remove or capture surface-level variations in sentence construction, label names, and synonymous connecting

terms, the final stage, information standardization, addresses surface-level variations from acronyms and synonymous phrases in the main concepts (e.g., '`financial experience`', vs. '`experience in finance`'). For some concepts, this is feasible by continuing the dissection process, standardizing the terms by dissection them to properties. Recognizing the meaning of acronyms and terms without repetitions of words, however, requires a level of external knowledge not directly known from the text. While humans can draw on background knowledge to recognize that a `BBA` is a bachelor level degree in business, or that the term `USA` and `United States` are the same, computers do not directly have this knowledge. This information can, however, be supplementally added, through a variety of approaches including manually specifying the degree and subject of terms such as '`BBA`', or by looking up the information on external databases to standardize '`USA`' and '`United States`' as the same location. As illustrated in Figure III.2, decomposing concepts to the sub-ontologies developed in Chapter 2 (and included in Appendix D), and standardizing each dimension, facilitates direct comparisons, irrespective of how the text is written.

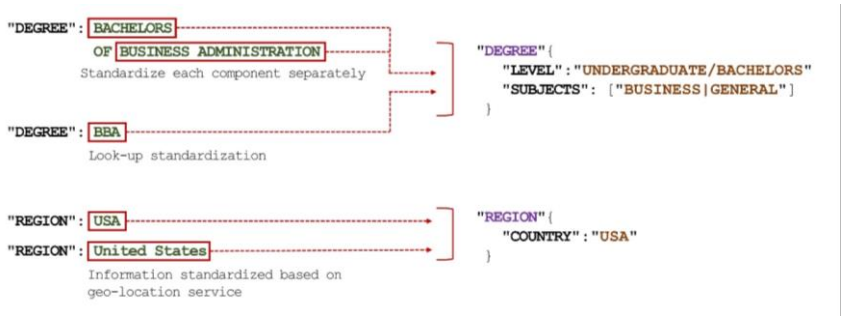


Figure III.2 Illustration of Standardization of Concepts

While additional consideration will be given to extend these approaches to represent entire sentences (which can be aggregated to represent entire managerial backgrounds), the three stages, i) concept identification, ii) semantic interpretation, and iii) standardization, provide the foundation for capturing representations of managerial backgrounds.

CHAPTER IV

Managerial Backgrounds Implementation Specifics

This chapter explains how the approach introduced in the previous chapter can be extended from individual relationships and simple sentences to populate the ontologies developed in Chapter 2, and illustrated in Figure IV.1 below.


```
{
  "ORIGINAL_SENTENCE":"Previously, Ms. Doe worked for Gorilla Software Inc., a NYSE-listed game developer,
                       as the Executive Vice President of Sales between July 2005 and June 2010, and before
                       that worked for Oculi Machined Parts as the Head of Marketing between January 2001
                       and April 2005.",
  "BACKGROUND":{"PERSON_NAME":{"ORIGINAL":"Ms. Doe","NAME_TITLE":"MS","LAST_NAME":"DOE"}
  "POSITIONS":[
   {
        "ORIGINAL":"worked for Gorilla Software, a NYSE-listed game developer, as the Executive Vice
                    President of Sales between July 2005 and June 2010",
        "JOB_TITLE":{"ORIGINAL":"Executive Vice President of Sales","LEVEL":"EVP","AREA":["SALES"]}
        "COMPANY_NAME":{"ORIGINAL":"Gorilla Software Inc.","NAME_CLEANED":"GORILLA SOFTWARE"}
        "START_DATE":{"ORIGINAL":"July 2005","YEAR":2005, "MONTH":7}
        "END_DATE":{"ORIGINAL":"June 2010","YEAR":2010, "MONTH":6}
   },
   {
        "ORIGINAL":"worked for Oculi Machined Parts as the Head of Marketing between January 2001 and April
                    2005",
        "JOB_TITLE":{"ORIGINAL":"Head of Marketing","LEVEL":"HEAD","AREA":["MARKETING"]}
        "COMPANY_NAME":{"ORIGINAL":"Oculi Machined Parts","ORIGINAL":"OCULI MACHINED PARTS"}
        "START_DATE":{"ORIGINAL":"January 2001","YEAR":2001, "MONTH":1}
        "END_DATE":{"ORIGINAL":"April 2010","YEAR":2005, "MONTH":4}
   }]
}
```


Figure IV.1. Illustration of Indicative Sentence Complexity and the Intent of the Process

While the process described in the previous chapter drew closely from information extraction approaches (e.g., Lample et al., 2016; Jurafsky and Martin, 2018), the relatively complex sentence structures in managerial backgrounds require separate consideration. This chapter begins by explaining the approach to reduce the sentence complexity so that the long and varied sentences can be populated to the ontologies. This is followed by more specific details

involved in transforming the sentences to the developed ontologies for each of the three stages: concept recognition, semantic interpretation, and standardization. Since the intention of this dissertation is to develop an approach that is sufficiently flexible to be capable of representing the breadth of managerial backgrounds, a large pool of managerial backgrounds were first collected. The managerial backgrounds used throughout the development process were sourced from the proxy statements of all US public firms over the period 2007-2017 (obtained directly from SEC/EDGAR).[29] This comprised of around 8 million sentences, including both directors and top managers, and the breadth of experiences, sentence structures, and acronyms included in these backgrounds helps ensure the flexibility and scalability of the population approach.[30]

Extension of the Information Extraction Approaches to Represent Entire Sentences

As indicated above, the primary way that representing entire sentences differs from the last chapter is the level of sentence complexity. While the prior approach is well suited to relationships between a small number of concepts, (ignoring sentences parts not specifically desired, e.g., Nakashole, Theobald, and Weikum, 2011; Biega, Kuzey, and Suchanek, 2013), the sentences in managerial backgrounds are typically long, often discussing multiple positions, experiences and qualifications. This is illustrated in Figure IV.2, showing how as the sentence length increases, the number of concept orderings substantially grows.

[29] Proxy statements were identified as by far the most consistent source of discussion on the backgrounds of executive officers and the board members.

[30] Specifically, the breadth and volume of material means that only modest adaptations are envisioned necessary to extend the approach to earlier and future time periods, and other communication mediums. Further details on how the material was extracted from the managerial backgrounds is included in Appendix B, and descriptive statistics of the material, derived through the development of the ontologies, are included in Appendix D.

```
Sentence: John Doe is the CEO
Concepts: PERSON_NAME IS MANAGEMENT_TITLE

Sentence: John Doe was appointed the CEO in 1997
Concepts: PERSON_NAME APPOINTED MANAGEMENT_TITLE IN DATE

Sentence: Previously, Ms. Doe worked for Gorilla Software Inc., a NYSE-listed game developer, as the
          Executive Vice President of Sales between July 2005 and June 2010, and before that worked for
          Oculi Machined Parts as the Head of Marketing between January 2001 and April 2005.
Concepts: TIME_BEFORE PERSON_NAME WORKED FOR COMPANY_NAME DETERMINANT COMPANY_DESCRIPTION AS
          MANAGEMENT_TITLE BETWEEN DATE AND DATE AND TIME_BEFORE WORKED FOR COMPANY_NAME AS MANAGEMENT_TITLE
          BETWEEN DATE AND DATE
```

Figure IV.2 Illustration of How Increasing the Sentence Length Increases Complexity

Since there are tens of thousands of unique concept orderings, it is unfeasible to directly write rules to populate the ontologies from concepts. The approach to allow the more complex sentence orders to be captured involves two important considerations, that while not ultimately impacting the populated ontologies developed in Chapter 2, substantially simplify the population process. These considerations are: i) capturing concepts at a relatively high level of granularity, and ii) temporarily grouping concepts before interpreting the meaning. Capturing concepts at a relatively high level of granularity involved avoiding unnecessarily complicating the semantic interpretation from too fine concepts. Specifically, while it is possible to represent '`BA in Engineering`' as `DEGREE IN ENGINEERING` or `DEGREE IN SUBJECT`, semantic interpretation is simplified if the term is represented as `DEGREE`, and then split into level and area in a second stage. Similarly, while '`Vice President of Manufacturing`' could be represented as `MANAGEMENT_TITLE OF MANUFACTURING` or `MANAGEMENT_TITLE OF AREA`, representing the whole term as `MANAGEMENT_TITLE`, and then separating it into parts, simplifies the semantic interpretation.[31]

[31] This level of concepts are used throughout the dissertation (including Figure IV.2 above) and are detailed further in Appendix C. While, as illustrated in Figure IV.2, this ordering is not simple, there is nevertheless less variation than if concepts had been split further.

There are, however, trade-offs from using higher level concepts. While capturing concepts at a high level reduces the complexity of interpreting the meaning, it also increases the number of terms in the concepts. For example, further aggregation such as combining management titles and companies (e.g., '`Vice President of Manufacturing of Gorilla Software`') or degrees and universities (e.g., '`Bachelor in Engineering from the University of Michigan`') means that the number of different terms substantially increases. While terms such as '`Vice President of Manufacturing`' are sufficiently general that they appear across managerial backgrounds, the aggregate of the management title and company name tend to be manager-specific. Temporarily grouping blocks of related concepts together achieves the benefits of reduced complexity, while avoiding increasing the number of terms in the concepts. As explained further below, concept orders such as `MANAGEMENT_TITLE OF COMPANY_NAME` can be temporarily grouped together, simplifying the overall ordering to facilitate semantic-interpretation. As illustrated in Figure IV.3, this enables long, complex concept orderings to be simplified, substantially reducing the overall number of unique orderings.

Figure IV.3 Illustration of Grouping Concepts

The remainder of this chapter will describe further considerations at each stage, including specifics of how the concepts were identified, greater consideration of the semantic interpretation, and the standardization and verification process.

Concept Recognition

The first stage of populating the ontologies is concept recognition. As noted previously, since concepts including `MANAGEMENT_TITLES`, `COMPANY_DESCRIPTION`, and `COMMITTEE` are not identified in standard packages (e.g., OpenNLP: Baldridge, 2005; NLTK: Bird and Loper, 2004; and Stanford NER: Finkel, Grenager, and Manning, 2005), as outlined in Figure IV.4, it was necessary to develop training data, allowing further terms in the concepts to be identified through machine learning.

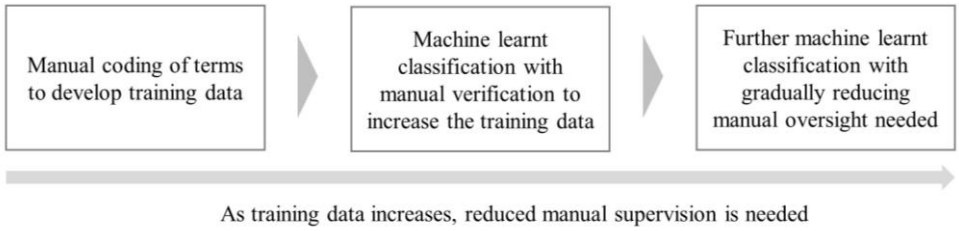


Figure IV.4. Summary of the Overall Process to Classify Terms into Concepts.

The initial training data was built up by manually tagging words to concepts, supplemented with various manually defined 'rules'. For example, capitalized words proceeding '`inc.`', '`corp`', and '`LLC`' were identified as likely company names and terms following '`degree in`' as likely degrees. As the number of fully classified sentences increased, machine learning was used, using neural networks (Specht, 1991) on the Tensor Flow platform (an open source machine learning platform released by Google: Abadi et al., 2016) to automatically classify words and phrases based on the training data.[32] This uses pattern recognition on the trained data to fit a model, allowing the terms in subsequent sentences to be classified.[33] While machine

[32] To avoid over-fitting, a dropout of 0.5 was used (Srivastava et al., 2014).

[33] For example, while the ability to recognize that `QUALIFICATION IN unknown_word` could likely be combined to a `QUALIFICATION`, based on classified data with a similar underlying pattern, the machine learning process is able to identify this and other groupings. In this way, new sentences, containing unknown words and phrases can be passed through the model, and each word or sequence of words, classified into a particular concept, allowing the process to be scaled with only manual oversight.

learning helped speed up the process of classifying terms, the entire process was heavily supervised to ensure the validity of the concepts, and avoid "semantic drift", a potential issue of recursive classification where the introduction of errors causes more errors to be introduced (Riloff and Jones, 1999; Curran, Murphy, and Scholz, 2007: 172). To ensure the accuracy of the terms in the concepts, and reduce the need for manual verification, a variety of automated and semi-automated approaches described below were used to verify the terms, and identify classification errors; the overall accuracy of the machine learnt classification of concepts is now in excess of 99%[34], with manual verification helping ensure that mistakes are identified.

Semantic Interpretation

The next stage in the process, semantic interpretation, involves populating the concepts to the ontologies. As noted, while there is substantial variation in the overall concept orders, this can be reduced by first grouping related concepts, and then interpreting the meaning from the ordering of the groupings. Figure IV.5 illustrates how by grouping related concepts, a single rule can interpret a wide variety of sentence orderings.

[34] That is, when run over the training data, in excess of 99% of the machine-learned classifications correspond to the expected concepts. While the training data currently comprises several million fully classed sentences, and using the full sample achieves the most accurate model, to allow relatively quick iterations and incorporations changes and corrections, typically the models are trained on a sub-sample of around 100,000 sentences. This achieves high classification rates on new sentences, while only taking several hours to re-train the model.

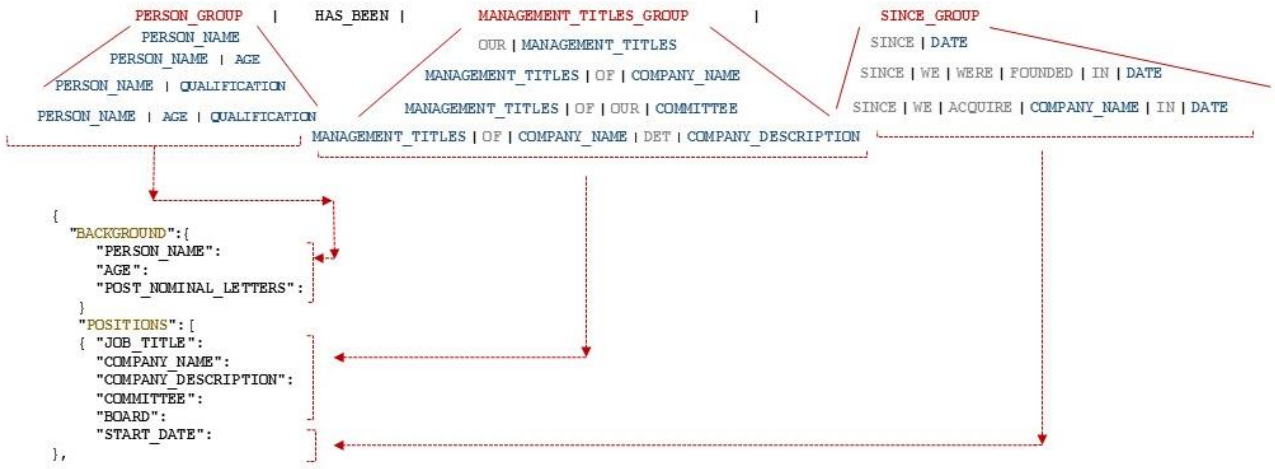


Figure IV.5. Overall Illustration of Groupings

These groupings were arrived at through consideration of similarity in the underlying material, grouping concepts that typically occur in similar places in the text, and populate the same parts of the ontology.[35] The approach to interpreting the overall meaning was then combined with more detailed semantic interpretation for each of the groupings. At this level, the task of associating concepts to appropriate parts of the ontologies is greatly simplified. Each of these used relatively simple rules, for example, that terms with the **PERSON_NAME** concept in the **PERSON_GROUP** should be associated with the **PERSON_NAME** property in the **BACKGROUND_DETAILS** primary ontology, or that the **DATA** concept in the **SINCE_GROUP** should be associated with the **START_DATE** property. This overall process is illustrated in Figure IV.6. While the number of groups increases with more complex sentences, the rate of growth is substantially less than at the concept level.

[35] As noted, these groupings are arrived at to facilitate implementation, and do not directly impact the populated ontologies. It would, for example, have been feasible to have grouped **MANAGEMENT_TITLE_GROUP** and the **SINCE_GROUP** together. While there would have been trade-offs from doing so (between a simpler overall sentence, but more varied orders within the grouping), it would not have ultimately impacted the populated ontology.

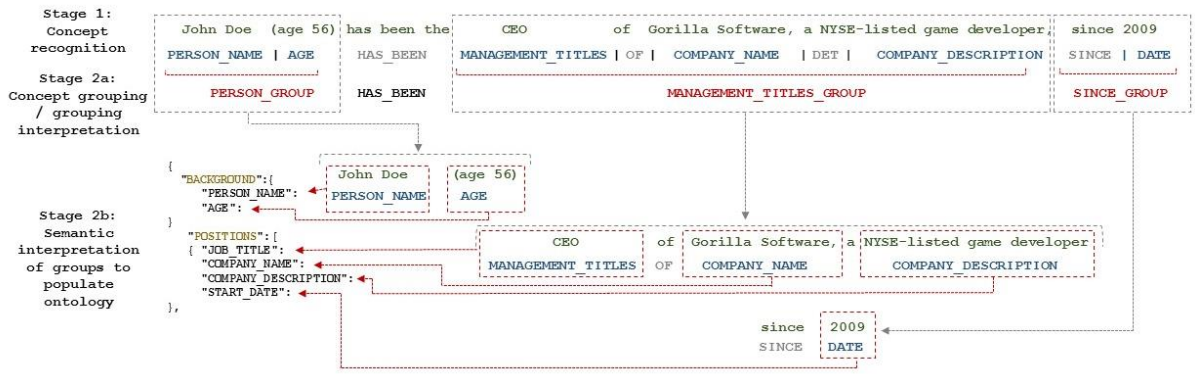


Figure IV.6. Example of Interpreting the Semantic Relationships for a Single Sentence

Standardization and Verification of Concepts

While the populated ontologies capture the properties of the text, to facilitate comparisons between sentences, the third stage involves standardizing the individual properties. The standardization approaches used included: i) standardization from dissecting terms to parts; ii) standardization from manual classifications; and iii) standardization from external databases. Standardizing by dissecting terms to parts parallels the process of standardizing the overall sentence, focused on capturing the specific properties of the terms. For example, some of the most commonly occurring properties in the **MANAGEMENT_TITLE** sub-ontology are **LEVEL**, **AREA**, **COUNTRY**. As illustrated in Figure IV.7 below, it is possible to break down management titles to these components, and populate the sub-ontologies in a similar manner as at the overall sentence level.

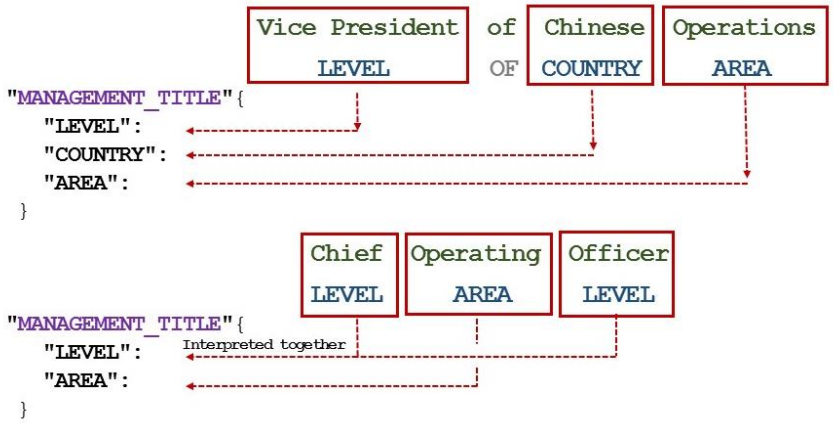


Figure IV.7 Illustration of Dissecting Concepts with Sub-Concepts

This process reduces the complexity; while there are around 50,000 unique management titles, the number of unique terms in each property of the sub-ontology is much less. For example, there are only around 2,000 unique levels, ranging from commonly occurring terms such as '`Vice President`', to less common permutations such as '`Acting Senior Vice President`'. Moreover, Figure IV.8 illustrates how by continuing the dissection process, splitting up the level property to its components, the 2,000 permutations can also be fully standardized.

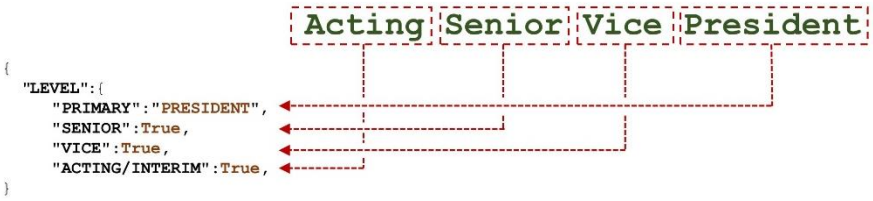


Figure IV.8 Illustration of Dissecting Sub-Concepts to Properties

As illustrated in Table IV.1 below, a large proportion of concepts can be dissected to the components of the sub-ontologies, with the number of unique terms reducing at each stage of the dissection process.

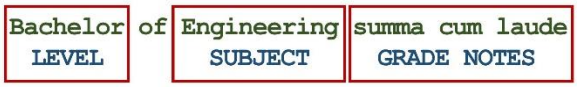
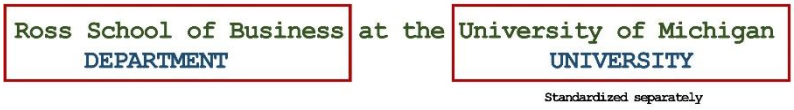
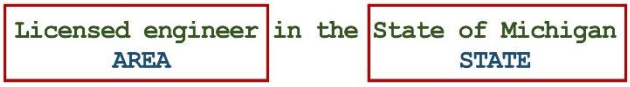
Concept	Example
QUALIFICATIONS	
PERSON_NAME	
EDUCATION_ INSTITUTION	
LENGTHS_OF_TIME	
DATES	
MANAGEMENT_TITLES	
COMMITTEE	
PROFESSIONAL_ LICENSE	
COMPANY_ DESCRIPTION	

Table IV.1 Summary of the Standardization Approaches

Once fully dissected, the number of terms at each property is typically small, amenable to manually standardizing any remaining variations. For example, as illustrated in Figure IV.10, in the terms '`over 5 years`' and '`more than five years`', both '`over`' and '`more than`' are treated synonymously, and '`5`' and '`five`' are both standardized to the number `5`, populating the sub-ontologies described in Appendix D.[36]

[36] Dissecting can also be extended to allow larger numbers to be interpreted from the sequencing; '`five hundred`', can be dissected to `QUANT HUNDRED`, which is then amenable to interpreting as `500`. Larger numbers, such as '`four million six thousand and twenty-five`' is also amenable to be interpreted in a similar manner.

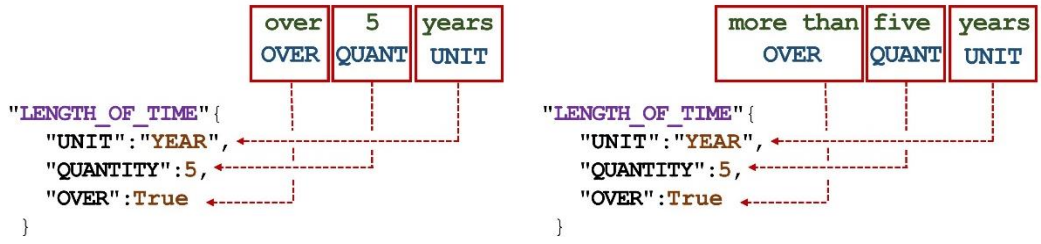


Figure IV.9 Illustration of Standardization Through Dissection

As well as making terms consistent, the additional benefit of dissecting the terms into components is that it helps ensure the validity of the terms in the concepts: although there are tens of thousands of terms at the concept level, making manual verification of each difficult, there are a much smaller number of terms in each of the properties. Manually reviewing the component terms, and checking the overall terms in the concepts conform to an expected sequencing of components (for example that management titles in the order `LEVEL OF AREA` or `LEVEL OF AREA AND AREA`, etc.) helps ensure that the terms are appropriately classified while identifying potential mistakes. This is illustrated in Figure IV.10, illustrating how by successively dissecting terms, the difficulty of verification is much simplified.

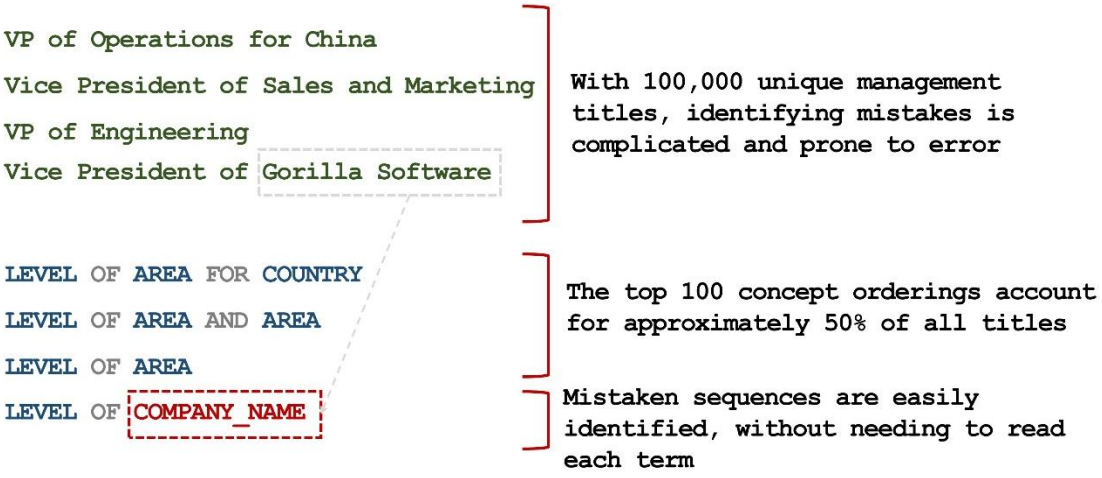


Figure IV.10 Illustration of Verification Through Dissection

As discussed further in Appendix E, while effort is still on-going to fully dissect terms, a high level of conformance is achieved across the majority of concepts (e.g., greater than 95% of terms fully dissected). This helps ensure i) that the ontologies are sufficiently flexible to capture the breadth of properties ii) the semantic rules and classification of terms to sub-properties are functioning as expected, iii) and the validity of the terms underlying the concepts. Moreover, while 95% is sufficiently high to illustrate the potential of dissecting terms to the sub-ontologies, as the classification process continues, much higher classifications rates are anticipated.

While the approach to dissect the terms to components works for the majority of concepts, it is less suited for i) terms are obscured by acronyms, and ii) concepts that lack the inherent repetition in the underlying words to facilitate standardization through dissection. For example, it is not feasible to ascertain from the label the level or area of '**CFO**', nor that '**Michigan**' is a state, while '**Canada**' is a country. Both of these pose challenges to interpreting the meaning, and require a level of understanding of the term not directly inferable from the text. Since there are only a relatively small number of acronyms in the context of managerial backgrounds (e.g., **COO**, **CTO**, **VP**, **EVP**, **SVP** and **BEng**, **Med**, **BBA**) it is feasible to manually specify the properties as appropriate. For example, for the term **CFO**, the level (**CHIEF-OFFICER**) and the area (**FINANCE**) are manually specified, and for the qualification **BEng**, the level (**UNDERGRADUATE/BACHELOR**) and the subject (**ENGINEERING**) are manually specified, allowing acronyms to be dealt with relatively easily.[37] There are however limits to manually specifying terms, and manually standardizing concepts such as **LOCATIONS** (e.g., '**Texas**'),

[37] As discussed further in Appendix D, while many qualifications such as BBA, BEng, MEd include the subject as part of the degree title (i.e., business, engineering, and education), other degree titles such as BA and PhD do not (i.e., they are not necessarily in art or philosophy); care was taken to only infer the subject appropriately. Moreover, care was taken to extend classifications of degrees beyond common degrees to include various university-specific idiosyncrasies, such as the bachelor level S.B. and A.B. degrees issued by Harvard.

EDUCATION_INSTITUTIONS (‘University of Michigan’), and COMPANY-INDUSTRY_AREAS (e.g., ‘manufacturer of application specific integrated circuits’) would be unfeasible.

To ensure that these terms are appropriately standardized, terms in the LOCATIONS, EDUCATION_INSTITUTIONS and COMPANY-INDUSTRY_AREAS concepts were each connected with external databases. As illustrated in Figure IV.11, for education institutions, standardization was based on searching the top result on a search engine (Yandex), and then connecting the domain of the first result to a standardized database (Hipo, 2015).[38] Similarly, the location details were standardized based on a geocoding service (currently Google Geocoding: Google, 2018).

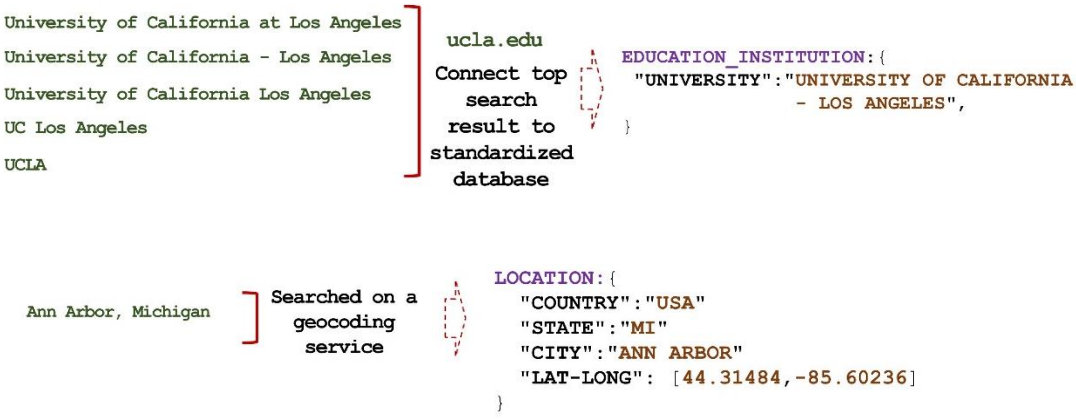


Figure IV.11 Standardization of EDUCATION_INSTITUTION and LOCATION Concepts

As well as helping to standardize the terms, this process also helps ensure the validity of the terms in the concepts; the domains of education institutions are typically .edu or .ac domain address, and domains not matching an associated university were flagged for manual review, as were locations that did not result in any matches on a geocoding service. Before describing the path for further standardization and validation, Table IV.2 summarizes the layers of validation

[38] In this instance, connecting terms via the search result has the slight advantage over ‘fuzzy matching’ (e.g., Zwick, Caristein, and Budescu, 1987), since it allows permutations that are very different, such as ‘UCLA’ and ‘University of California Los Angeles’ to be connected together.

that are incorporated throughout the process, each amenable to the large-scale textual analysis that help ensure that the ontologies reflect the underlying material and errors are identified.[39]

Approach	Description and Purpose
Qualitative development of ontologies	As described in Chapter 2, the ontologies were developed with substantial qualitative consideration of the underlying material. This ensures that the dimensions on which the text is dissected are relevant, and that the ontologies reflect the underlying text.
Validation by dissection	By dissecting concepts to underlying properties, and manually verifying the much reduced number of terms in the sub-concepts, and the sequencing of the sub-concepts, it is possible to validate a much larger number of terms. For example, the validity of concepts comprised of separate parts (e.g., `MANAGEMENT_TITLE`) can be assessed, despite there being tens of thousands of unique titles at the overall level.
Validation by external data-checks	By connecting terms to external databases, it is possible to verify concepts underpinned by a large number of labels, such as location information, that are unfeasible to manually verify, and lack the repetition in underlying words to allow dissection.
Validation by context	Checks that the context in which a term occurs in a sentence is appropriate; for example, while the concept sequencing `PERSON_NAME IS MANAGEMENT_TITLE AT COMPANY_NAME` is common, and appropriate, a concept sequencing such as `PERSON_NAME RECEIVED COMPANY_NAME FROM UNIVERSITY_NAME` is not common, and not likely to be correct (i.e., likely indicating that a degree acronym has incorrectly been classified as a company name). This validation includes three components: i. Manual checks to identify unlikely concept sequencing. ii. Machine-learned identification, where classifications through machine-learning are inconsistent with the classified concept. iii. Identification as concept sequencing that does not conform to that expected in the ontology This helps identify incorrectly classified concepts, irrespective of whether the concept has common terms (e.g., both concepts such as `LOCATION` and `MANAGEMENT_TITLE`), and in conjunction with validation through dissection and external checks, helps ensure the validity of concepts.
Manual oversight	Checks throughout the process to ensure that the ontologies are being populated in-line with those developed.
Face validity	Beyond documenting the ontologies, the examples, and summary statistics included in Appendix D illustrate that the ontologies, properties, and classifications have a high correspondence to what would be expected.

Table IV.2 Summary of Validation Approaches

The final standardization level is of the specific properties of the ontology. While at the specific property level the boundaries between classifications can be fuzzy (Lakoff, 1975; Zwick, Caristein, and Budescu, 1987; Murphy, 2002), as described further in Chapter V below, and in Appendix D, the intent is to reflect the dimensions of the underlying material, while allowing the

[39] While the ontologies are largely defined, and the process by which they are populated is largely functioning across managerial backgrounds with the vast majority terms classified appropriately, this validation is an on-going effort, and the current stage of development could be characterized as at a 'beta' stage (e.g., Abrahamsson et al., 2017), with the current status of verification reflected in Appendix E.

specific properties of the ontologies to be recoded, with properties combined as appropriate for a research question.[40] Full details of the levels captured is included in Appendix D. While the numbers of terms in many properties are sufficiently small to allow manual categorization, in the `COMPANY-INDUSTRY_AREA` property there remained a large number of terms describing the firm operations. Appendix E describes an approach, using surrounding discussion of the terms on Wikipedia, to enable terms not directly appearing in the NAICS classification manual to be connected to appropriate classification codes. This appendix also describes further developments intended to validate the classification of other fuzzy concepts, describing how crowdsourced verification (e.g., Kittur, Chi, and Suh, 2008) is intended to allow validation of the dimensions of the characterizations sub-ontology, and occurrence of subjects on department websites is intended to help validate groupings of subjects.

[40] For example, functional areas in the `MANAGEMENT_TITLE` concept including '`Finance`', '`Marketing`', '`Operations`' are distinct from one another, the functions have some underlying connections with '`Accounting`', '`Advertising`' and '`Manufacturing`' respectively, with research on functional areas aggregating the areas together (e.g., Michel and Hambrick, 1992; Ocasio and Kim, 1999). While these are each treated separately (along with other areas such as law and auditing), they can easily be combined as appropriate.

CHAPTER V

Dimensions of Interest, Aggregation, and Comparisons

This and the next chapter are intended to illustrate the theoretical opportunities of the developed approach. This chapter begins by systematically considering dimensions of the backgrounds most amenable for theoretical development, before giving more specific consideration to how dimensions can be captured and compared. While the discussion so far has illustrated how it is feasible to characterize individual sentences, since managerial backgrounds are comprised of multiple sentences, and the overall description of the top management team comprises multiple top managers, this discussion incorporates consideration of how the individual sentences can be aggregated. This is followed by examining how the ontologies facilitate consideration of the structuring of information. After providing the foundations to illustrate how the ontologies can be used to capture potentially interesting dimensions, the next chapter illustrates specific research questions facilitated.

Consideration of Theoretically Interesting Dimensions

While the ontologies capture substantially all the information in managerial backgrounds, their envisioned usage is forming constructs underpinned by specific, relevant, dimensions of the text. The research opportunities arise from the fundamental way in which the information departs from existing databases of employment histories; while existing databases seek to represent objective aspects of a manager's past, the approach developed here represents how that past is characterized. While certain aspects of managerial backgrounds are factual, such as dates of

employment or qualifications received (and manipulating these risks dismissal: Abrams, 2014), much of what is written, and especially how it is written, can be tailored. Specifically, since top managers will have accumulated varied skills across their careers, there is substantial opportunity to selectively draw on experiences, to highlight certain areas, while downplaying or omitting others. Similarly, it is possible to describe the organizations that managers have worked for in many ways, for example, highlighting success, pace of growth, and an international presence. The ontologies developed allow three fundamental forms of characterization to be explored: i) re-characterizations over time (e.g., Reger et al., 1994; Bolman and Deal, 2017), ii) different characterizations across mediums and audiences (e.g., Sutton and Callahan, 1987; Boje, 1991), and iii) deviations between characterizations and reality[41] (Meyer and Rowan, 1977; Westphal and Zajac, 2001).

By allowing direct comparisons between the dimensions of the text, the standardized representations facilitate comparisons of sentences, such as across-mediums or over time, even in the presence of surface-level variations. For example, as illustrated in Figure V.1, once the text has been represented in a standardized manner, sentences can be compared on each individual dimension of the text.[42]

[41] For example, omitting certain positions from positional history, or disproportionally discussing other experiences.
[42] The standardized structure of the ontologies make a variety of operations relatively easy (e.g., adding sentences, subtracting sentences, comparing orders within sentences etc.); the intention going forward is that functions will be made available to facilitate comparisons.

```
SENTENCE_1:{
  "ORIGINAL_SENTENCE":"From October 2001 to November 2004, she served as Vice President of Operations of QRS
                       Corp., a gold mining company
",
  "POSITIONS":[
   { "START_DATE":{"YEAR":2001,"MONTH":10},
     "END_DATE":{"YEAR":2004,"MONTH":11},
     "JOB_TITLES":{LEVEL":"VICE_PRESIDENT","AREA":["OPERATIONS"]}]
     "COMPANY":{"CLEANED":"QRS"}
     "COMPANY_DESCRIPTION":{"INDUSTRY":{"NAICS_3_DIGIT":212}}
}

SENTENCE_2:{
  "ORIGINAL_SENTENCE":"Between October 2001 and November 2004, she was previously the VP of Operations at QRS
                       Corp., a large international gold mining company.
",
  "POSITIONS":[
   { "START_DATE":{"YEAR":2001,"MONTH":10},
     "END_DATE":{"YEAR":2004,"MONTH":11},
     "JOB_TITLES":{LEVEL":"VICE_PRESIDENT","AREA":["OPERATIONS"]}]
     "COMPANY":{"CLEANED":"QRS"}
     "COMPANY_DESCRIPTION":{"INDUSTRY":{"NAICS_3_DIGIT":212},
                            "REGION":["INERNATIONAL"],
                            "CHARACTERIZATION":["SIZE_LARGE"]
}

DIFFERENCES_(2_minus_1):{
  "ADDITIONS":[
   { "COMPANY_DESCRIPTION":{"REGION":["INERNATIONAL"],
                            "CHARACTERIZATION":["SIZE_LARGE"]}
  "SUBTRACTIONS":{}
}
```


Figure V.1 Illustration of Position Comparisons[43]

In the above example, the subtle adjustment made to note that the company QRS Corp is large and international can be identified, despite the presence of surface-level differences (e.g., '`VP`' vs. '`Vice President`') that can complicate manually identifying changes. It is also possible to undertake comparisons, despite variations in the surrounding material or the order in which information is presented. This is important in facilitating comparisons between long communications, or different communication mediums, where differences in the order that information is presented are common. For example, the two texts in Figure V.2 below, describe a manager's background at two different companies (Gorilla Software and Oculi Parts); although

[43] To conserve space, and make the figures easier to compare, only relevant dimensions of the ontologies are displayed in this chapter.

the material is in a different order, and includes various other extraneous differences, by subtracting the description, it is possible to identify how the firms are characterized differently. While identifying such differences is not impossible by hand, it is non-trivial to consistently identify characterization differences between just two sentences; systematically comparing entire documents is almost impossible.

```
TEXT_1:{
  "ORIGINAL_TEXT":"Between October 2001 and November 2004, John was the head of marketing at Gorilla
                   Software Inc., and then between 2005 and 2009 he worked as the Vice President of Sales for
                   Oculi Parts Inc., an automotive parts manufacturer.
}

TEXT_2:{
  "ORIGINAL_TEXT":"Previously, John has worked as the VP of Sales for Ocili Parts, a leading manufacturer of
                   automotive components. Before that, he also worked for Gorilla Software, a large international
                   software development company."
}

DIFFERENCES:{[
  "COMPANY":"GORILLA_SOFTWARE",
  "ADDITIONS":[
    { "COMPANY_DESCRIPTION":{"REGION":["INERNATIONAL"],
                             "CHARACTERIZATION":["SIZE_LARGE"]}
                             "INDUSTRY":{"NAICS_3_DIGIT":511}}
  "SUBTRACTIONS":{}
,
  "COMPANY":"OCULI_PARTS",
  "ADDITIONS":[
    { "COMPANY_DESCRIPTION":{"CHARACTERIZATION":["LEADING"]}
  "SUBTRACTIONS":{}
]}
```

Figure V.2 Illustration of Position Comparisons Across Multiple Sentences

Mathematical Approaches to Compare Text

The above examples illustrate how once texts are in a standardized representation, operations with mathematical analogies are facilitated. Table V.1 describes in greater detail the foundational operations that allow comparisons between the standardized representations: i) simplification of sentences; ii) re-coding of dimensions, iii) additions, to aggregate sentences into blocks; iv) subtractions of sentences from one another to compare material; v) division of blocks

of text to allow relative comparisons of text.[44] Each of these operations can be implemented to enable different forms of comparisons; over time, across mediums; and comparisons in different audience responses to communications.

Operation	Description
Simplification	**Purpose**: Remove dimensions unnecessary for the research question from the ontologies While the ontologies are intended to represent substantively all of the discussion in the text, only certain parts are likely to be relevant for a particular question (in a similar manner to how only certain accounting measures are likely to be relevant to a particular research question). As such, an integral stage is determining desired dimensions, either ignoring, or removing dimensions superfluous to the research question. For example, dimensions such as the original text, or data of employment, are unlikely to be useful in many research questions, and removing them may make the ontologies easier to view and work with. While other properties, such as `PERSON_NAME` and `COMPANY_NAME`, may not provide the basis for comparison, they may be necessary for meaningful comparisons (e.g., seeing recharacterizations at specific companies).
Recoding (if necessary)	**Purpose**: Adjust any classifications based on the nature of the research question While the ontologies are intended to capture the dimensions of the text as noted earlier, at the very specific property level concepts can be 'fuzzy' (e.g., Murphy, 2002), and differences between concepts such as `MARKETING` and `ADVERTISING` may or not be desired. Thus, despite continuing effort to ensure a basis for classifications and validate the levels used, it is anticipated that a degree of re-coding may sometimes be required. The ontologies facilitate recoding of dimensions, with the ability to combine labels or further split concepts as desired. By providing the basis for a majority of dimensions, this task is much simplified; it is far easier to recode a limited number of functional-labels (such as designating that `MARKETING` and `ADVERTISING` should be combined) than to manually code the tens of thousands of permutations in the raw titles.
Additions (aggregation)	**Purpose**: Aggregate the properties as desired to summarize dimensions As described in Chapter 2, in the context of managerial backgrounds, it is appropriate to analyze each sentence separately, since in the vast majority of cases, the meaning of the sentence can be determined independently of surrounding discussion (i.e., limited use of pronouns such as '`this`' to refer to discussion in earlier sentences). Nevertheless, meaningful comparisons are likely to require comparisons at a higher level of analysis (e.g., the manager or top management team level). The nature of the aggregation is likely to depend on the question. At the highest

[44] While qualitative research inherently incorporates simplification and comparisons of aggregates of text, this typically relies on researchers' ability to synthesize the aggregate and differences (e.g., Glaser and Strauss, 1967; Eisenhardt, 1989), rather than the mathematical approach illustrated here. The more mathematical approach developed through the ontologies may have opportunities to complement more qualitative research; while qualitative research is well suited for considering the idiosyncratic information (that lacks the basis for directly represent in ontologies), the mathematical operations are more systematic, relying less on researchers' ability to synthesize the material, while facilitating very specific consideration of the ways in which the information is changing that is hard to capture even on a relatively small scale.

	level it would be possible to count a particular dimension, for example, the different forms of experiences that are discussed, or the breadth of industries, to allow a general characterization of a manager's profile (e.g., Crossland et al., 2014) as illustrated below.
	
	`SUMMARY_OF_FUNCTIONAL_AREAS:{`
	`"TECHNOLOGY":3,`
	`"OPERATIONS":2`
	`"MARKETING":1`
	`}`
	
	It is also possible to restrict aggregations to certain dimensions (i.e., crosstab-aggregation). For example, it is possible to aggregate company descriptions, restricted to specific companies, facilitating firm-specific comparisons.
	
	`SUMMARY_OF_BY_COMPANY:[{`
	`"COMPANY_NAME":"GORILLA SOFTWARE"`
	`"CHARACTERIZATION":{"SIZE_LARGE":2}`
	`},`
	`{`
	`"COMPANY_NAME":"OCULI_PARTS"`
	`"CHARACTERIZATION":{"LEADING":1}`
	`}]`
	
	Aggregation on particularly dimensions is particularly important when considering other communication mediums, where discussion on a particular topic may span multiple sentences.

Subtractions (comparison)	**Purpose**: Compare the nature of the material Subtractions are one way in which the material can be compared, with the ontologies facilitating comparisons on each dimension of the text separately (e.g., like-for-like comparisons restricted to the same company), as well as overall comparisons.
Divisions (relative comparisons)	**Purpose**: An alternative approach to compare the nature of the material While subtractions are one way in which separate blocks of material can be compared, an alternative is to consider the ratios of concept usage between mediums, allowing consideration of the relative differences of materials.

Table V.1 Mathematical Operations Feasible with Ontologies

While the above tables illustrate the primary ways in which the material can be compared, it is by no means restrictive. Specifically, the ontologies are essentially a multi-dimensional vector; by extracting a vector on a dimension of interest (e.g., a count of the functional areas that a manager has worked for), it is possible to make comparisons using measures common in management and strategy research, such as Euclidean distance/cosign similarities (e.g., O'Reilly III, Caldwell, and Barnett, 1989).

Figure IV.3, illustrates the process by which the ontologies allow relevant dimensions of the text to be captured, providing the basis for more traditional research analysis (e.g., regression analysis), where the captured dimensions can be dependent variables, independent variables, or controls.

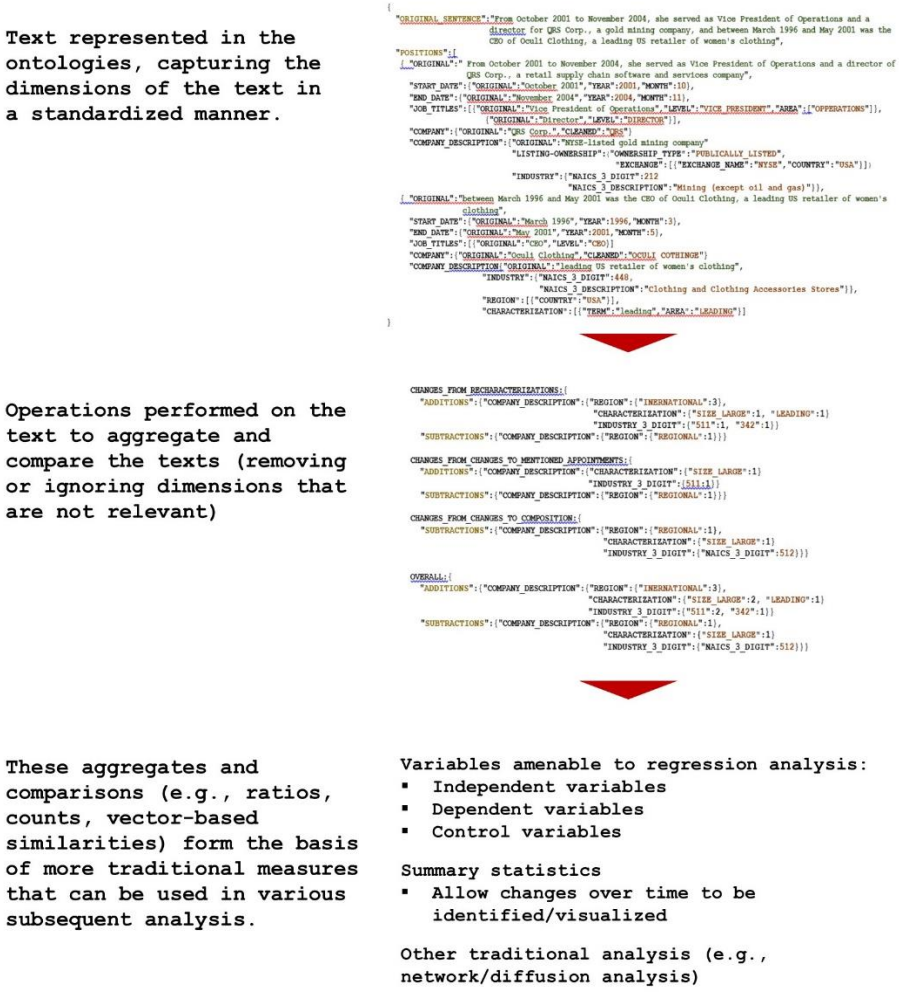


Figure V.3 Illustration of How the Process Integrates with Traditional Research

Facilitation of Like-for-Like Comparisons

One of the main advantages of having the information in the ontologies is that they allow like-for-like comparisons between texts, with the ability to see precisely the source of changes.

For example, changes to managerial backgrounds may arise from three qualitatively different reasons: i) the addition of new positions (such as a new outside direct position, or a change to the director's primary position), ii) recharacterizations of prior positions, iii) omission of prior positions. Being able to isolate the differences enables much more nuanced theorizing. An ability to identify systematic changes to how prior positions are characterized allows a level of theorizing beyond aggregate changes.

The ability to isolate the sources of differences may be particularly important when aggregating to the overall top management team; changes to the managerial profile resulting from appointing a new manager are qualitatively different to changes resulting from re-characterizing the presentation of existing managers' prior employment. Comparisons between the ontologies allow these differences to be easily separated. For example, Figure V.4 illustrates changes that can be identified by comparing the overall top management teams (either between years, or across mediums), indicating i) differences arising from recharacterizations of appointments; ii) differences arising from changes to the appointments mentioned for the same managers (e.g., new director roles, or complete omission of roles); and iii) changes arising from changes to the overall management composition. Since each of these are qualitatively different ways in which the positioning of the top management team can be adjusted, allowing the specific changes to be identified facilitates theorizing on the causes and consequences of the changes separately.[45]

[45] For example, appointing a manager with a significant financial background is a arguably more substantive change than adjusting a manager's prior experiences so that they are presented with more of a financial focus.

```
CHANGES_FROM_RECHARACTERIZATIONS:{
  "ADDITIONS":{"COMPANY_DESCRIPTION":{"REGION":{"INERNATIONAL":3},
                                      "CHARACTERIZATION":{"SIZE_LARGE":1, "LEADING":1}
                                      "INDUSTRY_3_DIGIT":{"511":1, "342":1}}
  "SUBTRACTIONS":{"COMPANY_DESCRIPTION":{"REGION":{"REGIONAL":1}}}

CHANGES_FROM_CHANGES_TO_MENTIONED_APPOINTMENTS:{
  "ADDITIONS":{"COMPANY_DESCRIPTION":{"CHARACTERIZATION":{"SIZE_LARGE":1}
                                     "INDUSTRY_3_DIGIT":{511:1}}
  "SUBTRACTIONS":{"COMPANY_DESCRIPTION":{"REGION":{"REGIONAL":1}}}

CHANGES_FROM_CHANGES_TO_COMPOSITION:{
  "SUBTRACTIONS":{"COMPANY_DESCRIPTION":{"REGION":{"REGIONAL":1},
                                        "CHARACTERIZATION":{"SIZE_LARGE":1}
                                        "INDUSTRY_3_DIGIT":{"NAICS_3_DIGIT":512}}}

OVERALL:{
  "ADDITIONS":{"COMPANY_DESCRIPTION":{"REGION":{"INERNATIONAL":3},
                                     "CHARACTERIZATION":{"SIZE_LARGE":2, "LEADING":1}
                                     "INDUSTRY_3_DIGIT":{"511":2, "342":1}}
  "SUBTRACTIONS":{"COMPANY_DESCRIPTION":{"REGION":{"REGIONAL":1},
                                        "CHARACTERIZATION":{"SIZE_LARGE":1}
                                        "INDUSTRY_3_DIGIT":{"NAICS_3_DIGIT":512}}}
```


Note: Since the ontologies for an entire management team are relatively long, to conserve space, an example of only the file result are displayed (these are however derivable from the underlying ontologies through aggregation/subtractions).

Figure V.4 Illustration of Comparison of Overall Top Management Team

Structuring of Information

While the discussion so far has focused on comparisons irrespective of order, the ontologies also provide the basis for considering information sequencing, an area that, with limited exceptions (Abbott, 1990; Kim and Jensen, 2011), has received little theoretical attention from strategy and management scholars beyond qualitative research (e.g., Martin et al., 1983; Boje, 1991).[46] While there are a large number of possible ways in which sequencing can be examined as described in Table V.2 below, the developed ontologies allow the sequencing of information to be captured and aggregated.[47]

[46] While sequencing is an important aspect of competitive dynamics, including work on first-mover advantage (e.g., Lieberman and Montgomery, 1988; Anderson and Tushman, 1990; Hambrick, Cho, and Chen, 1996), the focus of these areas is on temporal differences, rather than the broader impact of the order itself.

[47] While, as with structure more broadly (Fairclough, 1992), there are many possible ways in which order can be compared; the intentions is that it is possible to use the sequencing to capture constructs that have a close theoretical basis (e.g., Kim and Jensen, 2011; Harmon, 2018), with the aggregation and comparison approaches driven by theoretical considerations. Research areas that give substantial consideration to information ordering include

Operation	Description

Order extraction	**Purpose**: Extract ordering of dimensions for the ontologies
	
	The ontologies allow dimensions of text to be captured, with the ordering preserved within the ontologies (e.g., the order that positions, experiences and qualifications are described in the ontologies), and the order of discussion can be aggregated across sentences. It is possible to, for example, extract all functional areas discussed across managerial backgrounds, extracting the relevant information as appropriate from the **EXPERIENCES** and **POSITIONS** ontologies to generate a list of the sequence that terms are discussed, as illustrated below.
	
	Sequencing preserving sentence breaks[48]:
	`[["FINANCE","FINANCE"],["MANAGEMENT"],["ACCOUNTING"],["FINANCE"]]`
	
	Sequencing removing sentences breaks
	`["FINANCE","FINANCE","MANAGEMENT","ACCOUNTING","FINANCE"]`
	
	It is likewise possible to aggregate across the top management team, for example:
	`[M1:["FINANCE","FINANCE","MANAGEMENT","ACCOUNTING","FINANCE"],`
	`M2:["ACCOUNTING","ACCOUNTING","MANAGEMENT","ACCOUNTING","ACCOUNTING"],`
	`M3:["MARKETING","MARKETING","MANAGEMENT","FINANCE","FINANCE"],`
	`M4:["OPERATIONS","OPERATIONS","TECHNOLOGY","OPERATIONS"],`
	`M5:["FINANCE","FINANCE","ACCOUNTING","FINANCE"]]`

Table V.2 Approaches for Extracting the Order

The operations illustrated in the above table extend readily across other dimensions of discussion (e.g., the order that companies are listed in the text); just as it was possible to systematically identify slight changes to the presented information (e.g., across mediums or time), it is possible to use the ontologies to systematically identify changes to the ordering.

bioinformatics and DNA sequencing (Altschul et al., 1990; Edgar, 2010) and 'fuzzy' string matching approaches (Cohen, Ravikumar, and Fienberg, 2003). While these areas are unlikely to provide the theoretical basis for comparisons, their greater consideration to sequence comparison approaches may help operationalize desired constructs.

[48] Experience specific colors are added to facilitate connections to the theoretical discussion of sequencing in the next chapter.

CHAPTER VI

Large-Scale Theoretically-Centered Textual Research Opportunities

This chapter builds on the previous by illustrating specific research questions enabled by the ability to systematically capture standardized representations of managerial backgrounds. The discussion will be organized in two distinct levels: i) research questions at the individual manager level, and ii) research questions at the top management team level.

Individual Top Managers

While the impact of managerial backgrounds on decision making has received substantial academic attention (Hambrick and Mason, 1984; Cannella, Park, and Lee, 2008; Crossland et al., 2014), and significant research indicates that managerial backgrounds influence audience evaluations (Higgins and Gulati, 2003; Cohen and Dean, 2005; Chen, Hambrick, and Pollock, 2008; Zhang and Wiersema, 2009), background characteristics are typically taken as given, with little attention to the subtle ways in which firms may influence presentation. Consideration of how managerial backgrounds influence audience perceptions and decision making is focused on changes to the top management team's composition, such as appointing managers with specific backgrounds and demographic characteristics (e.g., Higgins and Gulati, 2003; Chen, Hambrick, and Pollock, 2008), or removing tainted managers (e.g., Arthaud-Day et al., 2006; Gomulya and Boeker, 2016), rather than on how managerial identity may be crafted without changes to composition. Research on impression management, however, illustrates how organizations can decouple perceptions from reality (Fiss and Hirsch, 2005; Westphal and Graebner, 2010; Rhee

and Fiss, 2014), and thus, rather than being an objective characterization of experiences (as typically assumed in upper echelon theory: Michel and Hambrick, 1992; Ocasio and Kim, 1999), managerial backgrounds can be considered a socially constructed phenomenon, with firms having opportunities to subtly tailor the presentation of experiences.

The ability to capture a representation of managerial backgrounds enables consideration of how the backgrounds are presented to external parties, and how these presentations impact back on the focal organization. Managerial backgrounds are well suited for analyzing impression management because they allow the 'content' to be analyzed independently of the way it is presented. While it can be difficult to separate deliberate impression management from more passive forms of communication, the ability to assess experiences independently of their presentation makes it possible to 'control' for the content. Specifically, firms may influence presentation, while still conveying the same underlying information, by changing the ordering (which can in turn influence what is considered important: McGraw, Lodge, and Stroh, 1990; Leung, 2014) or adjusting which experiences are elaborated on. It is also possible to examine changes to the presentation of information over time (e.g., re-characterization of prior experiences), or across mediums; since it is easier to use the same background across mediums and over time, such changes are likely to be purposeful. Moreover, since there exist archival databases of managerial career histories (e.g., BoardEx or ISS/RiskMetrics), it is possible to ascertain which experiences are being omitted or downplayed. The discussion below focuses on two specific opportunities for theoretical enrichment facilitated with the ontologies developed in this dissertation: i) how the presentation of managerial backgrounds may be tailored to specific audiences, and ii) how the personas created through the backgrounds may influence back on the firm.

Audience-Specific Impression Management

While substantial research indicates firms attempt to shape external perceptions (e.g., Elsbach, 2006; Pfarrer et al., 2008; Westphal and Graebner, 2010), and there is growing recognition that perceptions are audience-specific (e.g., Jensen, Kim, and Kim, 2012; Ertug et al., 2016), there is very little research systematically considering how managers and firms portray themselves differently to different audiences, beyond limited case study analysis (e.g., Sutton and Callahan, 1987; Boje, 1991). Research tends to take an audiences-perspective to explain perception differences, considering how heterogeneity in audience needs and expectations leads to variation in evaluations (Kovács and Sharkey, 2013; Jensen and Kim, 2014; Cattani, Ferriani, and Allison, 2014), rather than how audience-specific perceptions may be intentionally influenced by firms.[49] Given that most external audiences have limited direct managerial contact (Brown et al., 2015), adjusting how managers are presented across mediums may be particularly influential in shaping audience perceptions. By allowing variations in managerial positioning across different communication mediums to be examined, the ontologies provide an opportunity for researchers to consider audience-specific identity formation.

Specifically, the ontologies facilitate examining differences between the presentation of managerial experiences in corporate filings, which are focused on the investment community, and on company websites, which are more customer-focused, allowing subtle differences in presentations across mediums to be identified. As described in the previous chapter, by making it

[49] One exception to this would be work that considers how firms try and hide their identity to particular audiences for example through pseudonyms (Phillips and Kim, 2009) or contract brewers concealing the origins of their beer (Carroll and Swaminathan, 2000). Nevertheless, hiding one's identity from certain audiences is qualitatively different from the more subtle tailoring of a message for different audiences described here. Similarly, while institutional theory does give consideration to how firms may seek to satisfy different stakeholder demands, for example through ceremonial adoption of activities (Meyer and Rowan, 1977; Zajac and Westphal, 2004), there is less consideration as to how a firm may manage the impression of different audiences of the same activity (i.e., managerial backgrounds), and specifically how the language usage in different communication mediums are used to cater to audience-specific demands.

feasible to ‘subtract’ backgrounds from one another, the ontologies enable differences between backgrounds that would be hard to identify by hand to be systematically revealed; entire backgrounds can be substracted from one another, even in the presence of significant surface-level variations. Given that different organizational stakeholders have different needs and concerns, comparisons between settings enables consideration of how firms may engage in audience-specific impression management, highlighting different information to the two audiences, such as presenting the team’s management as having a technology background to customers, while a strong financial background to the financial community. Moreover, while this relatively sophisticated form of impression management may be interesting to examine in itself, it also provides the foundations for broader theorizing, such as considering the extent to which firms can maintain separate identities across mediums, and the broader consequences from attempting to do so.

Conformance to Personas

While the ontologies enable examination of how firms may influence the perceptions of external parties about a manager’s experiences, they also allow consideration of the influence of the created personas back on management. While a large body of research indicates that individuals conform to expected behaviors (e.g., Goffman, 1959; Philipsen, 1975; Hochschild, 1983; March and Olsen, 1984), how personas shape managerial action has received little theoretical consideration. Specifically, while substantial research in the upper-echelon tradition illustrates that managerial backgrounds influence decisions (e.g., Michel and Hambrick, 1992; Marcel, 2009; Crossland et al., 2014), the reasons why prior experiences influence decision making is less theorized; explanations still rest on the assumption that backgrounds influence

managerial attention and preferences (Hambrick and Mason, 1984; Hambrick, 2005).[50] Although the argument that experience influences attention and preferences is relatively taken-for-granted in the literature, it is not necessarily the entire explanation: rather the personas that managers create in presenting their backgrounds could influence audience expectations, and conformance to these role expectations (e.g., Merton, 1957), could contribute to the differences in managerial behavior, above and beyond the manager's direct experiences. The ability to measure the decoupling of conveyed managerial experiences, including what is emphasized or omitted, from the manager's actual experiences, provides opportunities to examine the extent to which behavior is influenced by conveyed personas, above and beyond prior experiences.

Individual Top Managers: General Discussion

The two areas are outlined in Figure VI.1 below, illustrating the possibility of dynamics in the impression management of managerial backgrounds, with presentation influencing audience expectations, which in turn may impact back on firm management. The envisioned expansion of the ontologies to other communication mediums, using the approach developed in the dissertation, would allow the broader evolution process to be examined. While the developed ontologies allow examination of impression management and the impact of the created personas on firm management, extending the ontologies to capture audience perceptions, for example through analyst reports, would allow the dynamics in the overall influence process to be explored.

[50] Writing in 2005, Hambrick noted that the way in which background filters attention "has not been studied as much as it needs to be; nor, to be honest, has [the process] ever been verified" (Hambrick, 2005: 114).

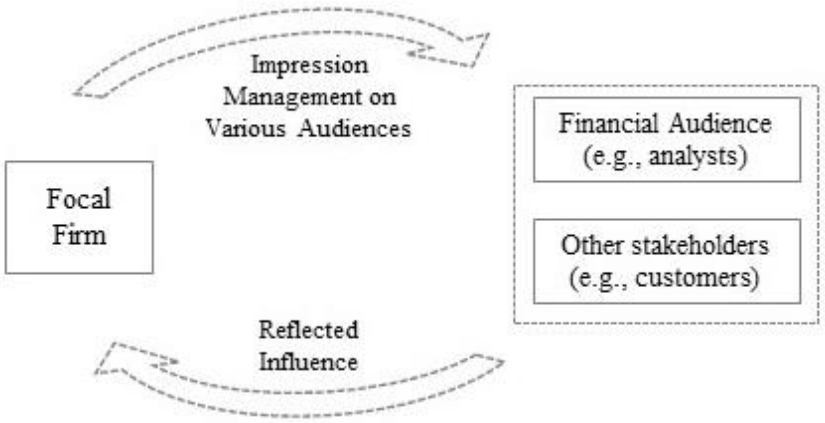
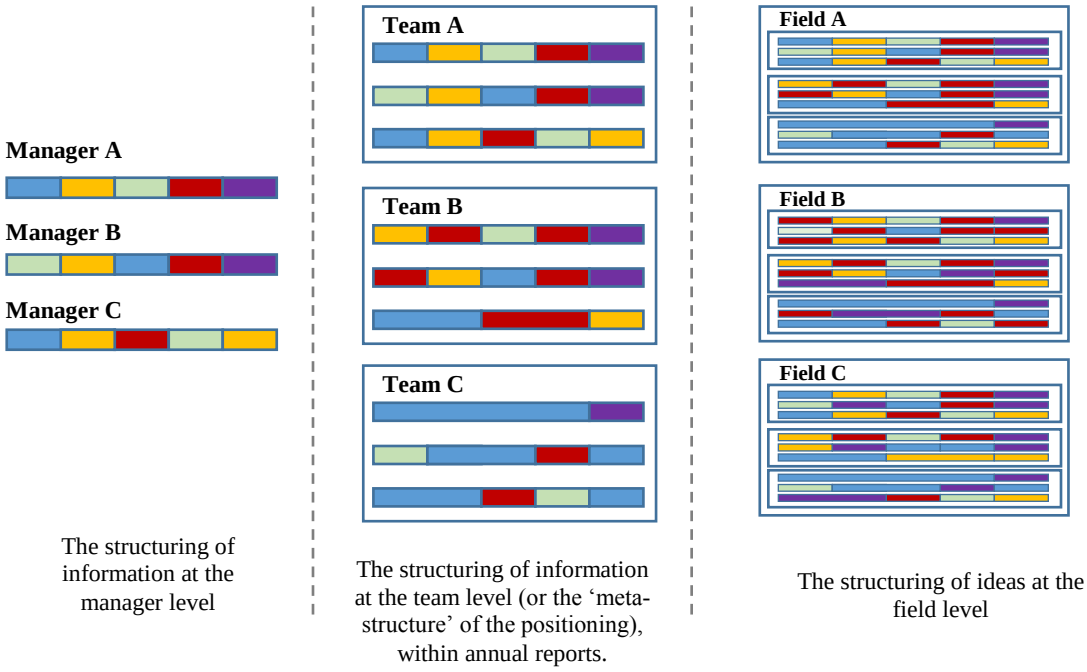


Figure VI.1 Impression Management Cycle

Overall Top Management Team

While the opportunities discussed so far consider just the positioning of an individual manager or CEO, the analysis can be extended to the group level. The relative simplicity and consistency of managerial backgrounds make them well suited to develop theory concerning how ideas are structured together, and how that structuring evolves. Specifically, managerial backgrounds offer a relatively clean way to capture and compare information ordering. Despite significant psychological research indicating that ordering fundamentally shapes information interpretation (e.g., McGraw, Lodge, and Stroh, 1990), with a limited number of exceptions (e.g., Kim and Jensen, 2011; Leung, 2014), little consideration has been given by strategy and management scholars to how information is ordered into an overall structure, and the impact of this ordering on audience perceptions. As noted in the previous chapter, there is very little systematic research on the influence of ordering, and being able to systematically capture the ordering that information is presented in documents is a necessary stage in furthering advancement of this area. As illustrated in Figure VI.2, the representations of the managerial backgrounds captured in the ontologies, allows consideration of the meta-structures at the top

management team and field level to be examined.[51] In addition to allowing the variations across communication mediums, and the evolution of structure to be explored, as discussed below, the ontologies enable consideration of two specific areas: i) the presentation of top management team diversity, and ii) the relative presentation of the board of directors and the top management team.



Note: The colors are used to represent a different dimension of experience.

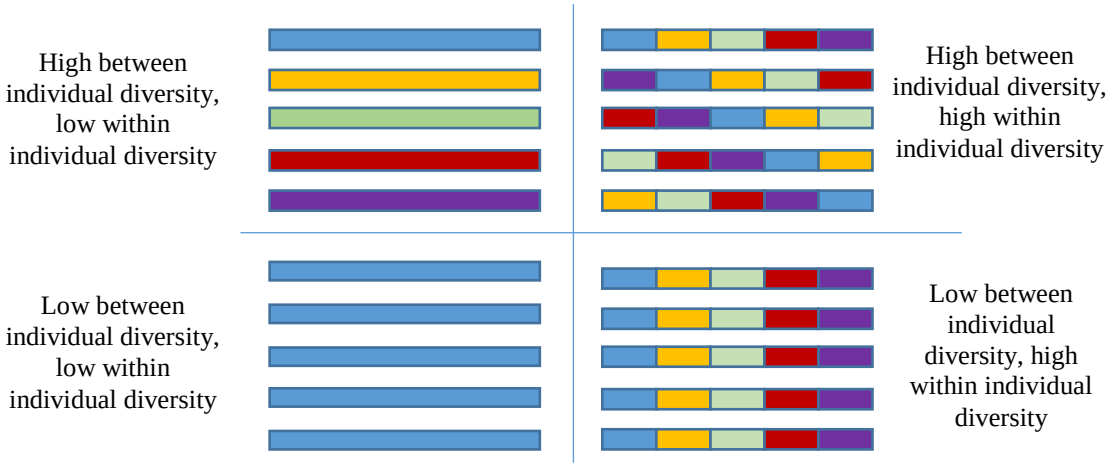
Figure VI.2. Illustrations of the Multiple Layers of Structure

Presentation of the Diversity of the Top Management Team

The first opportunity at the group-level is to build on research examining diversity in top management team experiences (e.g., Michel and Hambrick, 1992; Simons, Pelled, and Smith,

[51] As noted, once the high-level theoretical overview of the 'broad research opportunities' discussed has been presented for each communication medium, a more specific focused discussion will be given to domain-specific contributions. For example, not only do managerial backgrounds provide opportunities to enrich the literature on information ordering (e.g., Abbott, 1990), but they also provide opportunities to contribute to the top management team literature on the portrayal of diversity (e.g., Bernardi, Bean, and Weippert, 2002).

1999; Cannella, Park, and Lee, 2008), by examining how diversity in experiences are externally conveyed. Although there is research considering the presentation of the managerial team's demographic diversity (e.g., Bernardi, Bean, and Weippert, 2002), there is much less research considering how the broader types of diversity are presented. Considering the presentation of diversity is a theoretically interesting topic, partly because it can be theorized at different levels, including within individual diversity of experiences and between individual diversity, and also because there is no clear 'right' level of diversity (Hambrick, Cho, and Chen, 1996; Crossland et al., 2014). As such, there may be benefits to projecting different diversity depending on audience expectations, with the possibility for presentation to be tailored either based on the firm's context, or building on the discussion of audience-specific impression management above, based on specific audience expectations. The approach developed in this dissertation allows rich theorizing on the applicability of different diversity compositions, and the conditions under which firms try and project each of the four compositions illustrated in Figure VI.3.



Each horizontal bar corresponds to an individual manager, and each color illustrates a different type of experience.

Figure VI.3. Stylized Illustration of Different Forms of Managerial Diversity.

Presentation of Board Control

The final research opportunity using the developed ontologies is to examine the relative positioning of the top management team and the board of directors. Building on work indicating firms seek to influence perceptions of external control (i.e., Finkelstein and D'Aveni, 1994; Westphal and Graebner, 2010; Cohen, Frazzini, and Malloy, 2012), the ontologies make it possible to consider how the backgrounds of the top management team are presented relative to the board of directors. While research has examined ways that firms can indicate a separation of control between management and directors (e.g., Westphal and Graebner, 2010; Cohen, Frazzini, and Malloy, 2012), this work tends to focus on the portrayal of director independence. Control is however a relationship driven by relative power differences (Pfeffer, 1981), and in addition to the possibility that firms elevate the 'power' of external directors, firms may also create the perception of control by shifting how the experiences of internal directors are portrayed. The ability to capture the structuring of information in the ontologies allows identification of subtle changes to the presentation of internal and external managers, to increase the relative strength of external directors.

CHAPTER VII

Discussion

By developing an approach to transform textual information into a standardized representation, this dissertation illustrates how it is feasible to systematically capture meaning from text, enabling research that necessitates large volumes of rich, nuanced data. The approach developed is intended to allow an array of theoretically underpinned textual constructs to be directly measured, facilitating research on how meaning, the structuring of that meaning, and higher-level layers of meta-structures evolve. Moreover, the scalability of the approach allows the evolution process to be explored across entire populations of firms and extended time periods. While attempts in organizational theory to draw from computation analysis have to-date focused on latent analysis approaches (e.g., topic models: Magerman, Looy, and Song, 2010; Wilson and Joseph, 2015; Kaplan and Vakili, 2015), these approaches have inherent limitations in capturing nuanced, multi-dimensional constructs common in strategy and management theory. This dissertation draws from a fundamentally different computational approach, information extraction (e.g., Aggarwal and Zhai, 2012; Biega, Kuzey, and Suchanek, 2013), extending its scope to transform entire sentences into representations that capture the underlying meaning while preserving qualitative dimensions of what is said, allowing direct comparisons on specific dimensions of interest, and enabling easy aggregation to capture layers of textual structure.

The approach to systematically capturing a representation of meaning is developed in the context of managerial backgrounds. By detailing all stages in the approach, from the qualitative considerations of identifying the dimensions of the ontologies, to the computational process to

standardize the text into the ontologies, this dissertation provides a path for expansion into other communication mediums. The approach illustrates how representing the text in a limited number of primary ontologies (i.e., `BACKGROUND`, `POSITIONS`, `EXPERIENCES`, `QUALIFICATIONS`, and `PROFESSIONAL_LICENSES`) removes surface-level variation that has little to no impact on the underlying meaning, including acronyms, synonyms, slight differences in sentence constructions, and labels such as names. Throughout the development process, careful attention was given to capture or preserve abstract and subjective information, including characterization of past experiences, descriptions of the organizations that a manager has worked for, and the structuring of the information within and across sentences. The dissertation then shows how the standardized representations can be aggregated, at different levels of analysis, and compared on specific dimensions of interest.

The choice of managerial backgrounds for initial development also illustrates the potential of theoretically-centered textual analysis, with clear research opportunity using the developed approach. While research in the upper echelon tradition considers how the accumulated experiences can influence organizational leadership (Hambrick and Mason, 1984; Cannella Jr., Park, and Lee, 2008; Crossland et al., 2014), there has been little attempt to integrate impression management research (e.g., Bettman and Weitz, 1983; Elsbach, 2006; Fiss and Zajac, 2006) to explore subtle ways that the presentation of managerial experiences may be adjusted. Despite the importance of the perceptions of various external audiences to the ability of management to successfully lead their firms (e.g., Fredrickson, Hambrick, and Baumrin, 1988; Wiesenfeld, Wurthmann, and Hambrick, 2008; Westphal and Graebner, 2010), limited consideration has been given to the ways in which experiences may be recharacterized over time, or tailored to different audiences. By allowing managerial backgrounds to be systematically

represented, this dissertation makes it possible to identify differences in how specific experiences are described over time and across mediums. The representations also make it possible to examine multiple layers of information structuring, allowing consideration of how the structures of individual managerial backgrounds aggregate to meta-structures at the organizational level, and how these meta-structures evolve.

In a similar manner to how financial databases, board composition information, and the USPTO patent database have facilitated research where hand-collecting the data would have been unfeasible (e.g., Davis, 1991; Geletkanycz and Hambrick, 1997; Fleming and Sorenson, 2001; Carnabuci, Operti, and Kovács, 2016), systematically representing textual information across communication mediums can complement a wide range of scholarly inquiry. The nuanced characterization of the text in the ontologies allows scholars from diverse theoretical orientations to capture dimensions of theoretical interest, facilitating qualitatively different questions than possible with the more objective characteristics of firms, such as financial measures, patent counts, or existing board composition data.

While examing the organizational-stakeholder interface (e.g., Salancik and Meindl, 1984; Elsbach, 2006; Zott and Huy, 2007; Westphal and Graebner, 2010; Hiatt and Sangchan Park, 2013) may be the most direct opportunity from standardized representations of organizational and stakeholder communications, the approach has the potential to have a broader impact on the strategy and management field. Specifically, by allowing entire landscapes of communication to be characterized, the approach provides a qualitatively different setting to systematically examine the dynamics of firm interactions and how they adapt to environmental change. While the dynamics of how firms respond to the actions of competitors and adapt to their environment is a foundational pillar of strategy research, central to research on innovation (e.g., Bettis and

Hitt, 1995; Stuart and Podolny, 1996; Fleming and Sorenson, 2001), competitive dynamics (Barnett and Hansen, 1996), and evolutionary theories (Nelson and Winter, 1982; Levinthal, 1997), there are limited ways to capture representations of organizational landscapes. Researchers often generate simulated landscapes (e.g., Levinthal, 1997; Rivkin, 2000; Csaszar and Siggelkow, 2010), or focus on specific domains, such as the airline industry, (e.g., Baum and Korn, 1996; Hambrick, Cho, and Chen, 1996; Tsai, Su, and Chen, 2011) or patents (e.g., Stuart and Podolny, 1996; Carnabuci, Operti, and Kovács, 2016), where consistent data is systematically available across firms. Despite the abundance of available information, the ability of researchers to characterize the environment of public firms tends to be relatively coarse, proxying strategic positioning with financial indicators or the broad industries firms have operations in (e.g., Finkelstein and Hambrick, 1990; Geletkanycz and Hambrick, 1997; Crossland et al., 2014). An ability to systematically capture a rich representation of a communication landscape, and the dynamics of how firms are influenced by the communications of one another, provides a qualitatively different environment to explore and elaborate core strategy theories.[52] For example, drawing from research in the Carnegie tradition, it would be possible to characterize changes to organizational communications as a search process (e.g., Cyert and March, 1963; Levinthal, 1997), and to explore the dynamics of how firms tailor their communications to fit a changing multi-faceted environment. Moreover, as an increased number of communication types are systematically characterized (e.g., analyst reports or regulatory statements), it becomes possible to examine search dynamics across multiple distinct entity types.

[52] While studies to date do draw from textual information in analyzing competitive dynamics (e.g., Hambrick, Cho, and Chen, 1996; Boyd and Bresser, 2008), typically that research is focused on one variable, normally used to infer a strategic action; the approach developed in this dissertation allows a much richer characterization of the entire textual landscape, allowing greater consideration of the dynamics of how organizations present their operations (i.e., the competitive dynamics of organizational identity: Livengood and Reger, 2010).

This dissertation also attempts to re-center efforts to draw from computational linguistic advancements on theoretically relevant dimensions with direct connections to the text. While close theoretical connections between constructs and the underlying material has long been regarded as an important principle in social science research (e.g., Campbell and Stanley, 1963; Allen and Yen, 1979; Blackstone, 2012), the latent approaches that scholars have recently drawn from computer science (e.g., topic models: Magerman, Looy, and Song, 2010; Wilson and Joseph, 2015; Kaplan and Vakili, 2015) inherently have very weak connections between the constructs and theory. Even if it is possible to draw some forms of inference directly from data (e.g., Magerman, Looy, and Song, 2010; Grimmer and King, 2011; Bao and Datta, 2014), constructs with weak theoretical connections are unlikely to provide solid foundations for continued theoretical development. By allowing direct and transparent connections between the text and the information extracted, the developed ontologies help ensure that the relationship between constructs and the underlying material can be clearly made and conveyed, while allowing more nuanced theorizing than feasible with weakly connected measures.

Connections Between the Approach and High-Level Logics

By developing an approach to capture nuanced characterizations of text, this research also allows greater consideration of the levels of textual influence. While this dissertation takes the perspective that there is an important layer of meaning beyond merely individual words, captured only when considering the syntax of how those words are organized together (e.g., Matthews and Matthews, 1981; Van Valin and LaPolla, 1997), word usage still conveys a particular level of meaning (e.g., Hirsch, 1986; Abrahamson and Hambrick, 1997; Ocasio and Joseph, 2005; Loewenstein, Ocasio, and Jones, 2012). However, despite recognition of the different layers of meaning (e.g., Loewenstein, Ocasio, and Jones, 2012), there is very little

consideration of the conditions under which the layers are likely to impact; while qualitative research tends to emphasize the nuances in how meaning is constructed and conveyed (e.g., Gamson and Modigliani, 1989; Boje, 1991; Gioia and Chittipeddi, 1991; Martens, Jennings, and Jennings, 2007), more macro research tends to focus on higher-level logics (Ocasio and Joseph, 2005; Fiss and Zajac, 2006) with considerably less consideration of the dual influences. By systematically allowing consideration of both the nuance underlying the text, and the high-level logics,[53] this research offers the possibility for greater consideration of when the two layers of meaning are most influential. One specific opportunity to examine the different layers of influence may be to systematically consider differences between immediate and delayed responses to the release of organizational information. While an initial skim-read of released information is likely to reveal the high-level logics in the material, a more careful read is likely to reveal the more nuanced contextualization (Kintsch and van Dijk, 1978; Duggan and Payne, 2011), including explanation, justification, attribution, and discussion of mitigating factors (e.g., Scott and Lyman, 1968; Staw, McKechnie, and Puffer, 1983; Wade, Porac, and Pollock, 1997; Sonenshein, 2007).[54] Analyzing temporal differences in the market response to the release of organizational communications, between the immediate reaction where only skim-reads are feasible, and slightly delayed response once there is time for more careful read, may help reveal the impact of the contextualizations.

The ability to consider different layers of meaning also offers opportunities to examine attempts to simultaneously indicate conformity and distinctiveness. While typically considered opposing concepts (e.g., Finkelstein and Hambrick, 1990; Deephouse, 1999; Phillips and

[53] i.e., by aggregating without taking into the account the underlying nuance.
[54] While these contextualizations are uncommon in managerial backgrounds, they are common in other organizational communications, including annual reports and other stakeholder communications (e.g., Bettman and Weitz, 1983; Staw, McKechnie, and Puffer, 1983; Wade, Porac, and Pollock, 1997).

Zuckerman, 2001), the ability to assess the high-level logics independently of the lower nuance would allow consideration of how texts may be written to indicate conformity on an initial read while distinctiveness on more careful reads. Although there is growing awareness that individuals have inconsistent and contradictory preferences that may change over time (Tversky and Kahneman, 1981; Hoch and Loewenstein, 1991; Nordgren and Dijksterhuis, 2009), and that evaluation criteria often vary between initial screening and selection decisions (e.g., Gensch, 1987; Manrai and Andrews, 1998; Jensen and Roy, 2008), there is less consideration of the specific ways firms seek to meet changing evaluation criteria. The approach developed here allows for the exploration of attempts by firms to initially project overall conformity to expectations, while also indicating distinctiveness from other firms, with the salience of the two levels depending on engagement with the material

Development Process Going Forward

While this dissertation has developed the foundations to systematically capture representations of textual information, there are significant opportunities to develop on the approach. In addition to pursuing theoretical opportunities identified in the paper, future developments include: i) refinements to verify the concepts and ontologies, ii) elaboration of comparison approaches, iii) expansion into other forms of organizational communications, and iv) developments to allow qualitatively different forms of research questions to be explored. First, as explained in greater detail in Appendix E, as the ontologies become increasingly formed, the intention is to gradually expand the standardization and verification process to further identify any miscategorizations and validate the concepts. Part of this development includes expanding the automated verification so that a broader number of machine-learned terms can be automatically verified. This will then be supplemented with more qualitative

approaches including crowdsourced verification (e.g., Kittur, Chi, and Suh, 2008) of the CHARACTERISTICS sub-ontology, to help validate the dimensions, while facilitating its expansion to a broader range of characterizations (e.g., including negative characterizations of organizations by stakeholders).

The second intended development is to elaborate on the description of the approaches to aggregate and compare the ontology structures, illustrating the range of ways in which the ontologies can be used to capture constructs of interest. As well as describing the process by which the ontologies can be directly incorporated into research questions, specific consideration will be given to illustrate the ways that the ontologies facilitate theoretical examination beyond existing approaches (e.g., topic models: Magerman, Looy, and Song, 2010; Wilson and Joseph, 2015; Kaplan and Vakili, 2015). By illustrating the ability to directly capture a range of nuanced constructs from the text, this discussion is intended to complement the theoretical considerations in Chapter 1, illustrating how the ontologies facilitate examining questions that existing approaches are ill-suited to examine. Specific attention will be given to illustrate the ability of the ontologies to represent a level of meaning beyond merely the individual words, captured only when considering the syntax of how words are structured into sentences, and how sentences aggregate to meta-structures.

The third envisioned development is to extend the approach to represent a broad array of other textual information across the strategy, management, and social science fields. The approach was designed specifically to capture representations where there exists some form of deep-level similarity between texts (e.g., to a provide a basis on which to make the information consistent), allowing comparisons on desired dimensions despite the presence of surface-level variation (e.g., acronyms, synonyms, different label names, and slight differences in sentence

constructions). A large number of company and stakeholder communications have some deep-level similarities, and extending the range of source materials captured allows greater consideration of how communications evolve across mediums. For example, one relatively close extension is to capture the selection and evaluation criteria of management discussed by firms, allowing consideration of the reciprocal relationships between the presentation of managerial experience and the assessment decisions.[55]

The final envisioned development is to extend the approach to allow qualitatively different questions to be explored. As discussed in Appendix F, one envisioned development is to capture textual representations irrespective of the language used. Differences in the lexicon and the grammar structure (e.g., Stockwell, Bowen, and Martin, 1965), can both be considered surface variations, which the ontologies are already well suited to abstract, and developments in machine translations (Koehn et al., 2007; Wu et al., 2016) making it increasingly feasible to extend the approach to allow the multi-lingual evolutions of discussions to be explored.[56] Another potential extension is to capture discourse between two or more participants, for example, analyst conference calls. By considering who was on conference calls and exposed to a particular chain of questioning, the ability to systematically represent the information discussed makes it possible to use the approach developed, together with work on networks, to gain a nuanced understanding of how discourse evolves across the analyst network.

This dissertation is intended as one of the first pieces in a big picture. The start of a new journey, that while building on the past, is able to enrich strategy, management, and social

[55] An initial mockup, showing a possible way in which selection decisions could be represented, while also illustrating how the ontologies may be extended beyond managerial backgrounds, is included in Appendix F,
[56] Examples of how Spanish material could be represented without needing any changes to the developed ontologies are shown in Appendix F.

science theory, by allowing a new series of questions to be explored. Whether it takes five, ten or a hundred years to see the whole picture, I consider it an important future to be working towards.

APPENDICES

Appendix A

Text Colors Used in Dissertation

To make connections between the text and the figures easier to follow, a consistent color scheme, detailed in Table A.1 below, is used throughout this dissertation.[57]

	Description	Format	Format Example
Original 'raw' text	Example of unprocessed text, as it appears in the original source.	Colored dark green, in the original case.	John Doe is the CEO
Primary concepts	Used to represent the underlying type of the data.	Blue, upper case, with underscores between words	PERSON_NAME
Connecting concepts	Used to interpret the meaning, and how information populates the ontologies	Gray, upper case	FROM, TO, BETWEEN
Concept groupings	Used to group concepts together to reduce.	Red, upper case	MANAGEMENT_TITLES_GROUP
Primary ontologies	Used to signify the five primary data-structures	Dark yellow, uppercase, with underscores between words	BACKGROUND, POSITIONS, EXPERIENCES, QUALIFICATIONS, PROFESSIONAL_LICENSES
Sub-ontologies	Used for all data-structures other than the primary data-structures.	Purple, uppercase, with underscores between words	DATES, LENGTH_OF_TIME

Table A.1. Summary of Text Colors Used in the Dissertation

[57] Such colors are purely for visual purposes and are not necessary to follow the dissertation nor utilize the ultimate standardized representations of the text.

Appendix B

Collection of Managerial Background Data

As noted in the main text, the primary data source in this dissertation comprises approximately 8 million sentences extracted from proxy statements filed by US public firms from 2007-2017. All proxy statements were downloaded from EDGAR, and the sentences including the terms indicating reference to a manager (e.g., '`he`', '`she`', '`Mr.`', '`Mrs.`', '`Ms.`', '`Dr.`', as well as surrounding sentences) were extracted, with sentences not part of managerial backgrounds (for example discussing compensation decisions or performance targets), then filtered out. This data source serves as the bassis for the descriptive statistics included throughout the appendix.[58]

[58] It should be noted, that while the descriptive statistics are reflective of the underlying data, their primary purpose is to illustrate the face-validity of the extracted information; while they are indicative of the underlying material, because the process to refine the extraction approach is on-going, with refinements also being made to more accurately filter non-desired sentences, they should not be considered definitive.

Appendix C

Classifying the Text to Concepts

As described in the main text, to allow meaning to be interpreted from the text, the text is reduced to concepts. These concepts are split into two types, 'primary concepts' and 'connecting concepts'. The primary concepts form the basis of the information in the ontologies and account for much of the variation within the text. Connecting concepts are the semantic links that enable meaning to be interpreted.

Primary concepts

Table C.1 below includes details on all of the concepts that are classified, including an explanation of the concept, and example terms.

Concept	Explanation	Examples
QUALIFICATIONS	Full qualification (professional license are treated separately)	BBA, BEng in Mechanical Engineering, undergraduate degree in English, MBA
COMPANY_NAME	All company names, indulging company endings such as 'inc.', 'corp.' (note, education institutions are captured separately in a dedicated concept).	Gorilla Software Inc., Oculi Machined Parts Corp.
PERSON_NAME	All people names including name titles and suffix.	John Doe, Mr. John Doe, Mr. John Doe III, Jane, Mrs. Jane Doe, Dr Doe
EDUCATION_ INSTITUTION	All universities and colleges, including department names. While high schools are very rare in managerial backgrounds, this concepts could be expanded to include them.	University of Michigan, University of Michigan, Ross School of Business, MIT, Stanford University
LENGTHS_OF_TIME	All time periods, and any modifiers. While in the current data, this is largely comprised of years, the concept also incorporates months, weeks and days as appropriate.	5 years, in excess of 5 years, more than 5 years, around 5 years
DATES	Any dates or partial dates including modifiers to those dates (e.g., 'around). Date-like-events are also treated as dates, provided it is preceded by concepts including SINCE and TO[59]	2010, January 2010, January 1st 2010, 01/01/2010
MANAGEMENT_TITLES	All job titles, including modifier including	CEO, VP of Operations, Vice President of

[59] The approach for identifying date-like-events (i.e., descriptions of when something occurred, that take the place of dates) is slightly more complex than other dates. Specifically, while 'January 1st, 2010' can easily be identified as a date 'our founding' only indicates when something happened if preceded by words such as 'since' (e.g., 'since our founding' is treated as a date, while 'is one of our founding directors' is not).

	'founding'; typically capitalized.	Manufacturing, Director, VP of Sales and Marketing
COMMITTEE	All committees and sub-committees	audit committee, finance committee, nomination and compensation committee
BOARD	This concept capture all boards, in this context largely boards of directors	Board of Directors, Board of Governors, Board of Trustees
PROFESSIONAL_ LICENSE	All licenses, including the country or state that the license is valid in.	licensed engineer in the state of Michigan, CPA, chartered financial analyst
COMPANY_ DESCRIPTION	This concept captures any description of a company, including the operations, the country, public listing.	large multi-national company, leading producer of automotive components, startup tech company with operations
EXPERIENCES	Details on all experiences that a manager has, typically indicated by terms such as 'experience', 'background', 'insight into', 'knowledge of'	substantial insight into financial matters, experience in the pharmaceutical industry, wealth of experience with automotive companies

Table C.1 Summary of the Primary Concepts

Connecting concepts

While the terms that underpin the primary ontologies fit relatively directly within a particular concept, and the meaning of such terms tend not to vary by context (e.g., CEO almost refers to a management title), the meaning of many connecting words do vary depending on their usage. For example, the word '**from**' is used in managerial backgrounds to indicate when something happens from (e.g., '**from 2010**'), or university where someone got a qualification '**from**'. This makes it less feasible to group the terms into unambiguous concepts; while '**since**' sometimes has the same meaning as '**from**' (e.g., indicating when something began, such as '**from 2010**' vs. '**since 2010**') in other contexts it does not (e.g., '**since the University of Michigan**' vs. '**from the University of Michigan**'). Rather than attempt to force connecting terms together that sometimes have the same meaning, but other times do not, the approach taken was to only group terms that consistently shared the same meaning, while allowing multiple connecting concepts to serve as the same role when interpreting meaning. That is, in the ordering SINCE DATE PERSON_NAME HAS… and FROM DATE PERSON_NAME HAS… the two terms are treated synonymously, while in other contexts they may not.

Table C.2 shows the primary concepts that are used and examples of the usage.

Connecting concept	Example usage	Percentage of sentence concept occurs in
OF	PERSON_NAME IS MANAGEMENT_TITLE OF COMPANY_NAME	50.2
AND	PERSON_NAME SERVED AS MANAGEMENT_TITLE AND ON OUR BOARD	41.1
WHO (e.g., he, she, his, her)	WHO IS MANAGEMENT_TITLE	34.0
AS	PERSON_NAME SERVED AS MANAGEMENT_TITLE	33.4
DET (a, the)[60]	PERSON_NAME IS MANAGEMENT_TITLE FOR COMPANY_NAME DET COMPANY_AREA	30.5
SERVED	PERSON_NAME SERVED ON OUR BOARD	30.5
TO	FROM DATE TO DATE	29.5
FROM	i) FROM DATE TO DATE ii) PERSON_NAME RESIGN FROM COMPANY_NAME	25.3
IN	PERSON_NAME JOIN COMPANY_NAME IN DATE	24.9
HAS_BEEN	PERSON_NAME HAS_BEEN MANAGEMENT_TITLE	21.8
SINCE	SINCE DATE	16.2
WAS	PERSON_NAME WAS DET MANAGEMENT_TITLE	15.9
TIME_BEFORE (e.g, Before/Prior to)	TIME_BEFORE JOIN COMPANY_NAME	14.1
IS	PERSON_NAME IS MANAGEMENT_TITLE	10.8
OUR	PERSON_NAME SERVED ON OUR BOARD	10.6
FOR	PERSON_NAME WORKED FOR COMPANY_NAME FROM DATE TO DATE	10.3
ON	PERSON_NAME SERVED ON COMMITTEE	9.7
WITH	PERSON_NAME HAS BEEN MANAGEMENT_TITLE WITH COMPANY_NAME	9.2
AT	PERSON_NAME SERVED AT COMPANY_NAME	9.0
JOIN (e.g., Join/Joining)	PERSON_NAME JOIN COMPANY_NAME AS MANAGEMENT_TITLE	6.3
POSITION	PERSON_NAME POSITION INCLUDE	6.1
HELD	WHO HAS HELD WHO POSITION SINCE DATE	5.9
INCLUDE	PERSON_NAME POSITION INCLUDE	4.5
CURRENTLY	PERSON_NAME CURRENTLY SERVES AS	3.7
BRING-PROVDE	PERSON_NAME BRING-PROVIDE EXPERIENCE	3.7
COMPANY	PERSON_NAME HAS_BEEN MANAGEMENT_TITLE AT DET COMPANY	3.6
BY	PERSON_NAME WAS MANAGEMENT_TITLE UNTIL ACQUIRED BY COMPANY_NAME	3.5
WHERE	PERSON_NAME ATTENDED EDUCATION_INSTITUTION WHERE WHO RECEIVED QUALIFICATION	3.0
THROUGH	PERSON_NAME WAS MANAGEMENT_TITLE FROM DATE THROUGH DATE	2.8
APPOINTED	PERSON_NAME WAS APPOINTED MANAGEMET_TITLE IN DATE	2.4
RECEIVED	PERSON_NAME RECEIVED QUALIICATION	2.3
RETIRE	PERSON_NAME RETIRE IN DATE	2.2

Table C.2 Summary of the Connecting Concepts

[60] To simplify the orderings, determinants (e.g., '**a**', '**the**') are incorporated with surrounding connecting terms, such that '**with a**' is part of the WITH connecting concept. The connecting concept DET is thus only used when there is no other shrouding connecting concepts, such as COMPANY_NAME DET COMPANY_AREA.

While the above concepts account for the majority of interpretation, less common concepts are gradually being added to enable a broader spectrum of sentence constructions to be captured. For example, while the term ‘respectively’ only occurs in about 0.1% of sentences, it can impact the meaning of those sentences (e.g., `PERSON_NAME RECEIVED QUALIFICATION AND QUALIFICATION FROM EDUCATION_INSTITUTION AND EDUCATION_INSTITUTION RESPECTIVELY`).

Appendix D

Specifications of the Ontologies

This appendix is designed to fully specify each of the ontologies, including overall architecture of each of the ontologies, a description of key design considerations, and examples of how information is populated to the structure.

As discussed in the text, the overall approach to represent the meaning of the sentence is to dissect a sentence into simpler components and to dissect each element further until each of the dimensions of the text are captured and standardized. Specifically, as discussed in the main text, there are five primary information types (e.g., `BACKGROUND`, `POSITIONS`, `EXPERIENCES`, `QUALIFICATIONS`, and `PROFESSIONAL_LICENSES`), with essentially all of the discussion in a managerial background falling into one of these types. Each of these primary ontologies is then comprised of sub-ontologies (e.g., the sub-ontologies forming `POSITIONS` structure include `DATES, JOB_TITLES, COMPANY_DESCRIPTION,` etc.). This dissecting process is continued until every component is simplified to its most basic elements. For example, the text in `COMPANY_DESCRIPTION` such as '`a US manufacturer of automotive components traded on the NYSE`', is dissected further into the sub-structures `REGION_OF_OPERATION` (e.g., '`US`'), `AREA_OF_OPERATION` (e.g., '`manufacturer of automotive components`') and `LISTING-OWNERSHIP` (e.g., '`traded on the NYSE`'), with each of these sub-structures standardizing the individual text.

This section begins by documenting the most basic sub-ontologies that have no sub-dependencies (e.g., `DATES`, `LISTING-OWNERSHIP`). Next, sub-ontologies that have these dependencies are introduced (e.g., `COMPANY_DESCRIPTION`), and finally the overall the five primary-ontologies (`BACKGROUND`, `POSITIONS`, `EXPERIENCES`, `QUALIFICATIONS`, and `PROFESSIONAL_LICENSES`) are introduced. In this manner, the complexity of the ontologies is gradually built, and only after the more simple components are specified.

A summary of each of the data-structures, their dependencies, and the associated pages of the specification in this appendix is provided below in Table D.1.

Ontology	Dependencies	Component of	Page
DATE	-	POSITION, QUALIFICATION, COMPANY_DESCRIPTION	91
LENGTH_OF_TIME	-	EXPERIENCE	93
DEGREE	-	QUALIFICATION	94
EDUCATION_INSTITUTION	-	QUALIFICATION	98
LOCATION	-	PROFESSIONAL_LICENSE, MANAGEMENT_TITLE EXPERIENCE	100
FUNCTIONAL_AREA	-	MANAGEMENT_TITLE EXPERIENCE	101
MANAGEMENT_LEVEL	-	MANAGEMENT_TITLE	103
COMPANY-INDUSTRY_AREA		MANAGEMENT_TITLE, COMPANY_DESCRIPTION, EXPERIENCE	106
MANAGEMENT_TITLE	MANAGEMENT_LEVEL, FUNCTIONAL_AREA, LOCATION, COMPANY-INDUSTRY_AREA, CHARACTERIZATIONS	POSITION	108
COMMITTEE		POSITION	109
BOARD		POSITION	111
FIRM_FINANCIAL		COMPANY_DESCRIPTION	112
LISTING-OWNERSHIP		COMPANY_DESCRIPTION	113
CHARACTERIZATIONS		COMPANY_DESCRIPTION	115
COMPANY_DESCRIPTION	COMPANY-INDUSTRY_AREA; LOCATION; LISTING-OWNERSHIP, CHARACTERIZATIONS, FIRM_FINANCIAL, DATE	POSITION	117
COMPANY_NAME		POSITION	118
PERSON_NAME		BACKGROUND DETAILS	119
BACKGROUND_DETAILS	PERSON_NAME	OVERALL_ONTOLOGY	120
POSITION / POSITIONS	MANAGEMENT_TITLE, COMPANY_NAME, BOARD, COMMITTEE, DATE, COMPANY_DESCRIPTION	OVERALL_ONTOLOGY	121
EXPERIENCE / EXPERIENCES	LENGTH_OF_TIME, CHARACTERIZATION, FUNCTIONAL_AREA, COMPANY-INDUSTRY_AREA, REGION, LOCATION	OVERALL_ONTOLOGY	123
QUALIFICATION / QUALIFICATIONS	DEGREE, EDUCATION_INSTITUTION, DATE	OVERALL_ONTOLOGY	124
PROFESSIONAL_LICENSE / PROFESSIONAL_LICENSES	LOCATION	OVERALL_ONTOLOGY	126

Table D.1. Overall Summary of the Sub-Ontologies

Dates and Date-Like-Events

```
DATE:{
   "ORIGINAL": Original text                [String]
   "YEAR": Year, standardized               [Integer]
   "MONTH": Month, standardized             [Integer]
   "DAY": Day, standardized                 [Integer]
   "AROUND": Around/approximately           [True/False]
   "EVENT": Event that occurs on that day   [String: See Table D.2 below]
}
Used in: POSITION, QUALIFICATION
```

Figure D.1. Specification of the **DATE** Sub-Ontology

This sub-ontology is designed to capture a representation of dates (e.g., '**January 1st 2010**'), partial dates where the day/month may not be included (e.g., '**January 2010**' or '**2010**'), as well as date-like-events, which are used in the text in place of a specific date to signify when something occurred (e.g., '**[since] our founding**' or '**[until] he retired**', or in conjunction with a data/partial-date, such as '**[since] our founding in 2010**'). As illustrated in Figure D.1, this information is also standardized, with '**February**' is converted to the month '2', with date-like-events likewise standardized (e.g., '**since our founding**' and '**since we were founded**' are both treated as the firm's founding).

Value	Example
COMPANY_FOUNDING	[since] our founding
MERGER/ACQUISITIONS	[until] we acquired them in 1996 [until] 1996 when the company was acquired by Gorilla Software
DEPARTURE	[since] his retirement [until] he left the company in 2010 [following] his resignation
JOINING_FIRM	[since] she joined the firm in 2010
IPO	[Following] our IPO in 1996
RETIREMENT	[until] he retired in 2010

Table D.2. Summary of Date-Like Events

While the date-like-events currently captured are currently limited to events that occur in the context of managerial backgrounds, largely to indicate dates of employment (specifically, founding of the firm, acquisitions of other firms and retirements), the underlying representation can be readily be extended in the future to capture other date-like-events that occur in other company communications (e.g., '**We have supplied Gorilla Software *since the contract***

was signed in 2010'). Examples of the populated **DATE** sub-ontology are included in Figure D.2 below.

```
{
   "ORIGINAL":"1984",
   "YEAR":1984,
}
{
   "ORIGINAL":"around September 2006",
   "YEAR":2006,
   "MONTH":9,
   "AROUND":True,
}
{
   "ORIGINAL":"10/01/2006",
   "YEAR":2006,
   "MONTH":10,
   "DAY":1,
}
{
   "ORIGINAL":"his retirement in 2005",   # e.g. "[until] his retirement in 2005"
   "YEAR":2005,
   "EVENT":"RETIREMENT",
}
{
   "ORIGINAL":"our founding in 2010",     # e.g. "[since] our founding in 2010"
   "YEAR":2005,
   "EVENT":"COMPANY_FOUNDING",
}
```

Figure D.2. Examples of the **DATE** Sub-Ontology

Length of Time

```
LENGTH_OF_TIME:{
   "ORIGINAL": Original text                                                    [String]
   "UNIT": DECADE/YEAR/MONTH/WEEK/DAY                                           [String]
   "QUANTITY": Number of the units                                              [Decimal]
   "NON_SPECIFIC_QUANTITY": Textual description of the amount: SEVERAL, A_NUMBER_OF etc.   [Decimal]
   "OVER": Over/more than                                                       [True/False]
   "UNDER": Under/less than                                                     [True/False]
   "AROUND": Around/approximately                                               [True/False]
}

USED_IN: EXPERIENCES
```

Figure D.3. Specification of the **LENGTH_OF_TIME** Sub-Ontology

This sub-ontology captures a representation of how long something has occurred for (e.g., days, months, years, decades, etc.), as well as any modifications made to that description (e.g., '***approximately* a year**' or '***more than* three decades**' etc.). In the context of managerial backgrounds lengths of time are typically used to describe how much experience a manager has in a role, and although this is generally described in years, may also be in decades (which could in turn easily be converted into a different base). The sub-ontology is designed to be flexible to capture common modifiers (e.g., '**over**' or '**around**'), as well as imprecisely specified lengths (e.g., '**several years**' or '**a couple of years**').[61]

```
{
   "ORIGINAL":"over 30 years",
   "UNIT":"YEAR",
   "QUANTITY":30.0,
   "OVER":True
}
{
   "ORIGINAL":"more than 40 years",
   "UNIT":"YEAR",
   "QUANTITY":40.0,
   "OVER":True
}
{
   "ORIGINAL":"five and a half years"
   "UNIT":"YEAR"
   "QUANTITY":5.5,
}
{
   "ORIGINAL"several years"
   "UNIT":"YEAR"
   "NON_SPECIFIC_QUANTITY":"SEVERAL",
}
```

Figure D.4. Examples of the **LENGTH_OF_TIME** Sub-Ontology

[61] Although the sub-ontology does not make an assumption for what say constitutes 'several years', it would be easily possible to code subjective pieces of information, treating "*a couple* of years" as 2 years.

Degree

```
DEGREE:{
   "ORIGINAL": Original text                                                  [String]
   "DEGREE_LEVEL": Education level                                            [String - see Table D.3 below]
   "SUBJECTS": List of all subjects                                           [String - see Table D.4 below]
   "GRADE_NOTES": e.g., "summa cum laude"/"First class"/"distinction" etc.    [String[62]]
   "HONORARY": Honorary degree                                                [True/False]
}
USED IN: QUALIFICATION
```

Figure D.5. Specification of the DEGREE Sub-Ontology

This sub-ontology is designed to represent qualifications that individuals may receive, broken down and standardized into qualification level and subject. As would be expected, and illustrated in Table D.3, the vast majority of degree-level naturally fall into distinct levels: associated degrees ('associate degree'), undergraduate (e.g., 'undergraduate degree', 'bachelor', 'BA'), graduate (e.g., 'graduate degree', 'MA', 'Master's degree'), and post-graduate (e.g., 'PhD', 'Doctoral degree').[63] While the corresponding level could be easily ascertained for most degree acronyms based on the first letter of the qualification, with the letter 'B' signifying Bachelors (e.g., 'BA', 'BS', 'BBA', 'BEng' etc.) and 'M' signifying masters (e.g., 'MA', 'MEng', 'MEd'), care was taken to ensure that less common designations were associated with the appropriate degree level (e.g., the 'AB' and 'SB' designations, both Bachelor-level degrees issued by Harvard).

Level	Description	Examples	%
UNDERGRADUATE/BACHELOR	Most common qualification, typically a 4-year qualification.	Undergraduate degree, BEng, BA, B.B.A., Bachelor of Arts degree, Bachelor degree, AB, SB	53.4
GRADUATE/MASTERS	Graduate program, typically taken after a bachelors degree	Master's Degree, Graduate degree, MBA, MEng	30.3
DOCTORATE	Doctorate degrees	Doctorate, Ph.D., MD, Juris Doctor, Doctor of Law Degree	12.9
ASSOCIATE	A qualification below undergraduate level. In the US, typically 2-year degrees offered by community colleges.	Associate Degree, Associate of Science Degree, Associate of Arts Degree	0.2
UNSPECIFIED	Degree level is not specified.	degree, degree in mathematics, diploma, engineering degree	3.3

Table D.3 Summary of Qualification Levels

[62] In due course, it is anticipated that grade-notes will also be made consistent.
[63] While the split into the different levels appeared relatively direct (e.g., the vast majority of qualifications directly appeared in distinct categories of Associated, Bachelor, Masters, PhD), reference was also made to various guides discussing degree options (e.g., study.com), providing further validation of the four distinct degree levels.

The second component of the qualification is the subject. Typically the degree subject follows the degree level (e.g., '**a Bachelor's degree in Mathematics**', or a '**BA in Computer Science**'), however in some cases formed part of the degree name/acronym (e.g., an **M.B.A.** or a **MEng**, where the subjects, if not further specified, were taken as **BUSINESS|GENERAL** and **ENGINEERING** respectively).[64] Although most degrees typically have a single subject, there are cases where two or more subjects are listed (e.g., '**Bachelors of Arts Degree in mathematics and economics**'). In such cases, subjects are extracted and represented in the order in which that they occur.

Subject	Examples	%
BUSINESS-ACCOUNTING	Accounting, Accountancy	13.1
BUSINESS-GENERAL/UNSPECIFIED	Business, BBA, Business Administration	12.9
BUSINESS-FINANCE	Finance	8.2
BUSINESS-MANAGEMENT	Management Studies	3.9
BUSINESS-MARKETING	Marketing, Management and Marketing	1.7
BUSINESS-COMMERCE	Commerce	0.4
OTHER: BUSINESS-ADVERTISING, BUSINESS-BANKING, BUSINESS-ENTREPRENEURSHIP, BUSINESS-HUMAN_RESOURCE, BUSINESS-OPERATIONS, BUSINESS-ORGANIZATIONAL_BEHAVIOR, BUSINESS-STRATEGY		0.2
		40.4
MATHEMATICS	Mathematics, Applied Mathematics	2.0
COMPUTER_SCIENCE	Computer Science, Computer Sciences, Computer Systems	2.1
NATURAL_SCIENCE-CHEMISTRY	Chemistry, Organic Chemistry, Physical Chemistry	3.7
NATURAL_SCIENCE-BIOLOGY	Biology, Molecular Biology, Cell Biology	2.3
NATURAL_SCIENCE-PHYSICS	Physics, Applied Physics, Nuclear Physics	1.9
NATURAL_SCIENCE-GENERAL/UNSPECIFIED	Science, Sciences, Applied Science	1.8
NATURAL_SCIENCE-BIOCHEMISTRY	Biochemistry	1.1
NATURAL_SCIENCE-GEOLOGY	Geology	0.9
NATURAL_SCIENCE-MICROBIOLOGY	Microbiology	0.5

[64] Since many institutions grant degrees with formal titles that do not convey the actual subject (i.e., '**Bachelor of Arts**', '**Bachelor of Science**', and '**PhD**' are typically not in art, science or philosophy respectively), the subject was only taken from the degree name if that degree name was typically not followed by a separate subject. Specifically, for each degree title, a count was made of whether the degree was followed by a separately indicated subject (e.g., '**BA in mathematics**'); for degree titles including '**BA**', '**BS**', '**Bachelor of Arts**', '**PhD**', in line with expectation, the degree title typically was followed by a subject, (and typically that subject was not art or science). In other cases such as '**BBA**' or '**BEng**' it was less common to specify a subject separately (although naturally there were cases where the subject was included, such as '**BBA in business**', or where further specification was included, such as a '**BEng in electrical engineering**'). As such, in cases where the degree qualification indicates the subject (e.g., '**BBA**', '**BEng**', '**Bachelors of Engineering**', etc.), if no other subject is listed, then the subject is taken based on the degree title; in cases where the degree title typically does not convey the subject (e.g., '**BA**', '**BS**', '**Bachelor of Arts**', etc.), then if no other subject is listed, then the default specification is to leave the subject as unknown.

NATURAL_SCIENCE-BIOPHYSICS	Biophysics, Molecular Biophysics	0.2
NATURAL_SCIENCE-GENETICS	Genetics, Molecular Genetics	0.4
OTHER: NATURAL_SCIENCE-GEOPHYSICS, NATURAL_SCIENCE-GEOCHEMISTRY, NATURAL_SCIENCE-ECOLOGY		0.3
		13.1
ECONOMICS	Economics	10.9
LAW	Law, International Law, Business Law	0.4
ENGINEERING	Engineering, Civil Engineering, Industrial Engineering, Mechanical Engineering	18.2
PHARMACY	Pharmacy, Industrial Pharmacy	1.1
MEDICINE	Medical, Medicine, Surgery, Veteran Medicine	0.3
NURSING	Nursing, Psychiatric Nursing	0.2
IMMUNOLOGY	Immunology	0.2
HEALTH	Health, Public Health, Healthcare	0.1
		1.9
POLITICAL_SCIENCE	Political Science	2.1
HISTORY	History, Art History, European History	1.5
PSYCHOLOGY	Psychology, Industrial Psychology	1.1
LANGUAGES	English, French, Spanish, German, Russian	1.0
EDUCATION	Education, Physical Education, Elementary Education	0.6
COMMUNICATIONS	Communications, Communication, Organizational Communication	0.6
INFORMATION_SCIENCE	Information Systems, Computer Information Systems	0.3
LITERATURE	Literature, English Literature, Comparative Literature	0.3
OTHER: ZOOLOGY, CRIMINOLOGY, SOCIOLOGY, LIBERAL_ARTS, MUSIC, JOURNALISM, CLASSICS, THEOLOGY/RELIGION, PHILOSOPHY, HUMANITIES, URBAN_PLANNING, ARCHITECTURE, ANTHROPOLOGY, PHYSIOLOGY, GEOGRAPHY, SOCIAL_WORK		1.6
		9.1
ALL OTHER SUBJECTS		2.2

Table D.4 Summary of Qualification Subjects

Examples of degrees standardized by subject and levels are included in Figure D.6 below.


```
{
   "ORIGINAL":"B.S. degree, magna cum laude, in accounting",
   "DEGREE_LEVEL":"UNDERGRADUATE/BACHELORS",
   "SUBJECTS":["BUSINESS-ACCOUNTING"]
   "GRADE_NOTES":"magna cum laude"
}
{
   "ORIGINAL":"Bachelor of Arts degree in Music and History",
   "DEGREE_LEVEL":"UNDERGRADUATE/BACHELORS",
   "SUBJECTS":["MUSIC","HISTORY"]
}
{
   "ORIGINAL":"M.B.A.",
   "DEGREE_LEVEL":"GRADUATE/MASTERS",
   "SUBJECTS":["BUSINESS-GENERAL/UNSPECIFIED"]
}
```

Figure D.6. Examples of the **DEGREE** Sub-Ontology

A slight difficulty in capturing the degree information occurs when the degree is 'split' between a university, for example, '**received a Bachelors from the University of Michigan in Engineering**' or '**has an undergraduate degree from MIT in mathematics**'. The solution currently being adopted[65] for this is to create a concept for subject only, such that this standardizes as **RECEIVED QUALIFICATION FROM EDUCATION_INSTITUTION IN SUBJECT**., and then to interpret the qualification and subject together, combing **QUALIFICATION IN SUBJECT** 'on the fly' to reveal the qualification (as discussed in the main text though, in general it is generally easier to interpret the meaning from semantic ordering without such modifiers, so in general this approach is avoided).


```
{
   "ORIGINAL":"Bachelor [from the University of Michigan] in Engineering",
   "DEGREE_LEVEL":"UNDERGRADUATE/BACHELORS",
   "SUBJECTS":["ENGINEERING"]
}

{
   "ORIGINAL":"PhD degree [from Stanford University] in Engineering",
   "DEGREE_LEVEL":"DOCTORATE",
   "SUBJECTS":["ENGINEERING"]
}
```


Figure D.7. Example of a Split-Degree in the **DEGREE** Sub-Ontology

[65] This has yet to be fully implemented, but will in due course.

Education Institution

```
EDUCATION_INSTITUITON:{
   'ORIGINAL': Original text                          [String]
   'UNIVERSITY': University name (standardized)       [String]
   'DEPARTMENT': Department name                      [String[66]]
}

    USED_IN: QUALIFICATION
```

Figure D.8. Specification of the `EDUCATION_INSTITUTION` Sub-Ontology

This sub-ontology is designed to capture institutions, as commonly included in managerial backgrounds to describe the institution issuing the qualification. As illustrated in Figure D.9, this may also include the department in the university. Since there are multiple possible ways of writing the same university name, the university name is also included in a standardized format (i.e., represented the same, irrespective of who the information is entered). This is achieved by first searching the name of the universities on the internet[67]. For each result, the domain of the first search result was taken (e.g., '`UCLA`', '`University of California Los Angeles`' , '`University of California at Los Angeles`' all had the top search result ucla.edu), and then this domain was connected to the associated university in a database of university domain addresses (Hipo, 2015). Table Figure D.5 illustrates face validity of connecting universities to the standardized name; as well as illustrating the broad range of permutations standardized, the universities that managers most commonly received their qualifications from correspond with what would be expected: large, prestigious, US universities.

[66] It is anticipated that in due course department names will also be standardized, making partial department names such as '`Ross Business School [University of Michigan]`', consistent to '`Stephen M. Ross School of Business`', while also categorizing the department area as '`BUSINESS`'.

[67] The English version of the search engine Yandex was used to allow automated searches; if the top search result was a 'generic' result (e.g., a directory of universities), the next result was used (or manually reviewed/corrected as appropriate). As well as ultimately allowing the information to be subsequently integrated with other information, such as the location of the University (information also contained in the database connected to), the approach also had the advantage over 'fuzzy string matching' approaches (see Cohen, Ravikumar, and Fienberg, 2003) by allowing terms that share very few character, and hence are considered very different on fuzzy matching approaches (e.g., '`UCLA`' vs. `University of California Los Angeles`') to be associated together.

Standardized Education Institution	Examples	%
STANFORD UNIVERSITY	Stanford; Stanford University	4.5
UNIVERSITY OF CALIFORNIA, BERKELEY	University of California , Berkeley; University of California at Berkeley	2.1
HARVARD UNIVERSITY	Harvard; Havard University; Harvard University , Massachusetts	2.1
UNIVERSITY OF TEXAS AT AUSTIN	University of Texas at Austin University of Texas , Austin University of Texas	1.9
UNIVERSITY OF MICHIGAN	University of Michigan; University of Michigan , Ann Arbor; University of Michigan at Ann Arbor	1.8
UNIVERSITY OF CHICAGO	University of Chicago	1.8
YALE UNIVERSITY	Yale Yale University	1.7
UNIVERSITY OF CALIFORNIA, LOS ANGELES	University of California at Los Angeles University of California , Los Angeles UCLA	1.6
CORNELL UNIVERSITY	Cornell University, Conrell	1.6
MASSACHUSETTS INSTITUTE OF TECHNOLOGY	M.I.T Massachusetts Institute of Technology	1.5
PRINCETON UNIVERSITY	Princeton University	1.3
UNIVERSITY OF SOUTHERN CALIFORNIA	University of Southern California	1.3
UNIVERSITY OF ILLINOIS	University of Illinois	1.3
UNIVERSITY OF VIRGINIA	University of Virginia	1.3
UNIVERSITY OF WISCONSIN	University of Wisconsin	1.1
DARTMOUTH COLLEGE	Dartmouth College,	1.1
DUKE UNIVERSITY	Duke University	1.1
PENNSYLVANIA STATE UNIVERSITY	Penn State University Pennsylvania State University	0.9
BOSTON UNIVERSITY	Boston University	0.9
UNIVERSITY OF NOTRE DAME	University of Notre Dame	0.9
OTHER (1236 Standardized Institutions)		68.0

Table D.5 Most Common Standardized Universities

Examples of the populated ontologies are shown in Figure D.9 below.


```
{
   "ORIGINAL":"University of Michigan, Ross School of Business",
   "UNIVERSITY":"UNIVERSITY OF MICHIGAN",
   "DEPARTMENT":"Ross School of Business"
}
{
   "ORIGINAL":"Harvard",
   "UNIVERSITY":"HARVARD UNIVERSITY"
}
{
   "ORIGINAL":"Stanford University"s Law School",
   "UNIVERSITY":"STANFORD UNIVERSITY"
   "DEPARTMENT":"Law School"
}
```


Figure D.9. Examples of the **EDUCATION_INSTITUITON** Sub-Ontology

Location

```
LOCATION:{
   "ORIGINAL": Original text                                          [String]
   "COUNTRY": Country                                                 [String]
   "STATE": US State / Canadian Province                              [String]
   "CITY": City                                                       [String]
   "LAT-LONG": GPS coordinates (based on the center of the region)    [List of two floats]
}
USED_IN: PROFESSIONAL_LICENSE, MANAGEMENT_TITLE, EXPERIENCE
```

Figure D.10. Specification of the `LOCATION` Sub-Ontology

This structure is designed to represent specific location information discussed in the text (e.g., in the countries that a manager discusses having experience in, or the location of companies that they have worked for). To facilitate comparisons, this structure standardizes all of the location details, with supplemental information populated as appropriate (e.g., if a US state is listed, then to add the country as USA), and is also supplemented with latitude and longitude information based on the center of the region (this information is currently sourced from Google Geocoding: Google, 2018). Although naturally there are cities that share the same name (e.g., London in the United Kingdom vs. London in Ontario, Canada), typically either one city has a much larger population than the other, or sufficient disambiguation information is already included (e.g., while '`Washington`' may be ambiguous, it is typically written as '`Washington DC`' or '`Washington State`'); going forward, it is intended that there will be fine-grained options to adjust how ambiguous cases are handled.

```
{
   "ORIGINAL":"United States of America"
   "COUNTRY":"USA"
   "LAT-LONG":[37.09024,95.71289]
}

{
   "ORIGINAL":"Michigan"
   "COUNTRY":"USA"
   "STATE":"MI"
   "LAT-LONG":[44.314844,-85.60236]
}

{
   "ORIGINAL":"Ann Arbor, Michigan"
   "COUNTRY":"USA"
   "STATE":"MI"
   "CITY":"ANN ARBOR"
   "LAT-LONG":[44.31484,-85.60236]
}
```

Figure D.11. Examples of the `LOCATION` Sub-Ontology

Functional Area

```
FUNCTIONAL_AREA:{
   "ORIGINAL": Original text                    [String]
   "PRIMARY_AREA": Original text                [String see Figure D.6]
   "SUB_AREA": Original text                    [String]
}
USED_IN: MANAGEMENT_TITLE, EXPERIENCE
```

Figure D.12. Specification of the **FUNCTIONAL_AREA** Sub-Ontology

This sub-ontology is designed to capture functional areas, used in context of managerial backgrounds as part of managerial titles and experiences.[68] As noted in the text, while permutations in label names are removed as part of the standardization (e.g., **finance** vs. **financial**), the primary areas broadly reflect the labels in the underlying material, with recoding allowing areas such as **ADVERTISING** and **MARKETING** to be combined if desired, and less frequently occurring functional areas to be collapsed into an **OTHER** category (e.g., Ocasio and Kim, 1999). The sub-area reflects any permutations on the primary (e.g., '**consumer marketing**'), with the ontology expandable to allow these modifications to ultimately also be standardized (such permutations are not common among management titles, they are more common when describing experiences).

Functional area	Examples of job titles including functional area	%
FINANCE	Chief Financial Officer; Vice President of Finance; Vice President, Finance; Senior Vice President, Finance; Senior Vice President of Finance, CFO	45.5
OPERATIONS	Senior Vice President of Operations; Vice President, Operations; Chief Operations Officer; Director of Operations; Executive Vice President of Operations, COO	8.2
MARKETING	Chief Marketing Officer; Vice President of Marketing; Director of Marketing; Vice President, Marketing; Marketing Director	6.2

[68] When classifying managerial titles and experiences, a distinction is made between functional areas, and industry area. This is based on the generality of the experiences, and specifically whether the area transcends industries (a distinction that is reflective of existing research on functional areas: Michel and Hambrick, 1992; Ocasio and Kim, 1999). For example, areas such as finance, marketing, operations, legal transcend organizations; financial experiences for example is important to businesses in areas far from financial-focused institutions, and a majority of firms have a Chief Financial Officer. This is in comparisons to pharmaceutical experiences, which are most pertinent to the pharmaceutical industry. While a more systematic quantitative approach is intended to supplement the qualitative split (based on how specific areas are to particular industry segments), it is not envisioned that will result in any substantial differences from the qualitative considerations used currently.

SALES	Vice President of Sales; Vice President, Sales; Director of Sales; Sales Manager; Senior Vice President of Sales	4.4
HUMAN_RESOURCES	Vice President of Human Resources; Senior Vice President, Human Resources; Director of Human Resources; Chief Human Resources Officer; Human Resources Manager	3.4
ADMINISTRATION	Chief Administrative Officer; Vice President of Administration; Chief Financial and Administrative Officer; Chief Administration Officer	3.1
RESEARCH/TECHNOLOGY	Director of Research; Vice President of Research; Research Director	2.5
INVESTMENTS	Chief Investment Officer; Senior Investment Officer; Global Chief Investment Officer; Director of Investments	2.3
STRATEGY	Chief Strategy Officer; Senior Vice President , Strategy	1.8
COMPLIANCE	Chief Compliance Officer; Compliance Officer; Corporate Compliance Officer	1.8
ENGINEERING	Vice President of Engineering; Director of Engineering; Vice President, Engineering; Engineering Manager	1.5
BANKING	Chief Banking Officer; Head of Investment Banking; Head of Commercial Banking	1.5
AUDITING	Audit Manager; Senior Audit Manager; Head of Internal Audit	1.4
RELATIONS	Director of Investor Relations; Vice President of Investor Relations; Manager of Media Relations	1.1

Table D.6 Summary of Managerial Functional Areas

```
{
   "ORIGINAL":"Finance"                    # e.g., Chief Finance Officer
   "PRIMARY_AREA":"FINACNE",
}
{
   "ORIGINAL":"Consumer Marketing"  # e.g., Head of Consumer Marketing
   "PRIMARY_AREA":"MARKETING",
   "SUB_AREA":"Consumer"
}
```

Figure D.13. Examples of the FUNCTIONAL_AREA Sub-Ontology

The ontologies can also handle the acronyms, such as CFO or COO. For such acronyms, the 'original' term also appear in the MANAGEMENT_LEVEL sub-ontology, with the standardized level CHIEF-OFFICER.

```
{
   "ORIGINAL":"CFO"
   "PRIMARY_AREA":"FINACNE",
}
{
   "ORIGINAL":"COO"
   "PRIMARY_AREA":"OPERATIONS",
}
```

Figure D.14. Examples of acronyms in the FUNCTIONAL_AREA Sub-Ontology

Management Level

```
MANAGEMENT_LEVEL:{
   "ORIGINAL": Original text                                                    [String]
   "PRIMARY_LEVEL": Primary seniority level of manager, e.g., 'President       [String, See Table D.7 below]
   "ASSISTANT":                                                                 [True/False]
   "DEPUTY":e.g., Deputy Vice President                                         [True/False]
   "VICE": e.g., Vice President / VP                                            [True/False]
   "SENIOR":e.g., Senior Vice President                                         [True/False]
   "EXECUTIVE":e.g., Executive Vice President[69]                               [True/False]
   "LEAD": e.g., Lead Director                                                  [True/False]
   "DIVISIONAL": e.g., Divisional manager                                       [True/False]
   "GENERAL": e.g., General Manager                                             [True/False]
   "CO-LEVEL":e.g., CO-CEO                                                      [True/False]
   "FOUNDING": e.g., Founding Manager                                           [True/False]
   "INDEPENDENT": e.g., Independent Director[70]                                [True/False]
   "NON-INDEPENDENT": e.g., Non-Independent Director                            [True/False]
   "ACTING/INTERIM": e.g., Acting CEO                                           [True/False]
   OTHER: ACTIVE, VISTING, TENURED, DESIGNATED, HONORARY, INCUMBENT,
          CLINICAL, ADJUNCT, PRESIDING, EMERITUS, SOLE
   "CLEANED":  Representing desired dimensions of title[71]                     [String]
}
USED_IN: MANAGEMENT_TITLE
```

Figure D.15. Specification of the **MANAGEMENT_LEVEL** Sub-Ontology

This sub-ontology is a component of **MANAGEMENT_TITLE**, designed to capture the components of a manager's title that are associated with the seniority level (with functional areas, industry areas, and regions represented separately). In order to represent the approximately 2,000 different permutations on the manager's title, the sub-ontology is comprised of two parts: the overall designation (e.g., '**President**' or '**Director**'), and then a series of modifiers (e.g., '**Vice**', '**Senior**', '**Deputy**'), that cover the breadth of possible adjustments that can be made to the primary title. While an individual job title can have only one primary level, it is possible to have multiple modifiers.

---

[69] The term Executive can occur in two usages, either as a modifier or the main job title. When '**executive**' occurs as part of a job title, such as '**Executive Vice President**', it is treated as a modifier; when it occurs independently e.g., '**[is an] Executive**', it is treated as the primary title.

[70] While it is unfeasible to be both independent and non-independent simultaneously, they are treated separately, taking the value **True** if directly if the title directly is stated as independent / non-independent and omitted if not stated (this is consistent with other modifiers).

[71] Since many modifications to a job title are unlikely to be unnecessary for particular research questions, the intention is that a 'cleaned' version of the title could also be returned, standardizing the level to a single string that represents the desired dimensions, while suppressing non-desired dimensions. At present, the default is set to incorporate the main job permutations (specifically Executive, Vice, Senior, Deputy and General), such that both '**Acting Co-Vice President**' and '**Vice President**' standardizing as **VICE_PRESIDENT**. The dimensions that are included/suppressed will ultimately be able to be specified; whether or not a particular dimension is relevant or not is research question dependent.

PRIMARY_LEVEL	Examples (each of which may be modified, e.g., Vice/Senior/Lead)	%
PRESIDENT	President	25.6
DIRECTOR	Director	21.6
CHIEF-OFFICER	Chief Technology Officer, Chief Financial Officer, CFO, CTO, COO	9.7
CEO	CEO, Chief Executive Officer, Chief Executive	9.3
MEMBER (e.g., of the committee)	Member	9.1
CHAIR[72]	Chairman, Chair, Chairperson, Chairwoman, Chaired	7.8
PARTNER	Partner	2.3
MANAGER	Manager	2.1
MANGING_DIRECTOR	Managing Director	1.7
SECRETARY	Secretary	1.6
TREASURER	Treasurer	1.2
COUNSEL	General Counsel	1.1
EXECUTIVE	Executive	0.9
TRUSTEE	Trustee	0.6
HEAD	Head, Headed	0.5
OTHER/UNCLASSIFIED	PROFESSOR; COMMISSIONER; GOVERNOR; DEAN; FELLOW; LECTURER; CHANCELLOR; PROVOST; EDITOR	4.9

Table D.7. Primary Managerial Levels

Examples of the sub-ontologies are shown in Figure D.20 below. As illustrated, for acronyms and where the title is 'split' with a functional area between, the (e.g., '**Chief Technology Officer**'), the entire term is included in the original field.


```
{
   "ORIGINAL":"Interim Chief Technology Officer"
   "LEVEL":"CHIEF-OFFICER",
   "INTERIM":True,
   "CLEANED":"CHIEF-OFFICER"
}

{
   "ORIGINAL":"CO-CEO"
   "LEVEL":"CEO",
   "CO-LEVEL":True,
   "CLEANED":"CEO"
}

{
   "ORIGINAL":"Vice President"
   "LEVEL":"PRESIDENT",
   "Vice":True,
   "CLEANED":"VICE PRESIDENT"
}
```


[72] While the gendered terms '**Chairman**' and '**Chairwoman**' are by default grouped together with '**Chair**', and '**Chairperson**', reflecting the underlying role, the intention, as part of a broader future development to allow gendered language to be examined across organizational communications (e.g., Kendall and Tannen, 1997), is to have an option to also split out these terms.

```
{
   "ORIGINAL":"Senior Vice President"
   "LEVEL":"PRESIDENT",
   "Vice":True,
   "Senior":True,
   "CLEANED":"SENIOR VICE PRESIDENT"
}
```


Figure D.16. Examples of the **MANAGEMENT_LEVEL** Sub-Ontology

Company-Industry Area

```
{
   "ORIGINAL": Original text                                                    [String]
   "NAICS_3_DIGIT": The associated 3-digit NAICS-code                           [String]
   "NAICS_3_DESCRIPTION": Description of the associated 3-digit NAICS-code      [String]
   "ENTITY_TYPE": General characterization of the entity in the text            [String - See Table D.8 below]
   "MULTIPLE": Whether notes 'several', or plural such as 'companies'[73]       [True/False]
}
USED_IN: COMPANY_DESCRIPTION, MANAGEMENT_TITLE, EXPERIENCE
```

Figure D.17. Specification of the COMPANY-INDUSTRY_AREA Sub-Ontology

This sub-ontology captures a representation of the industry. As described separately in Appendix E, the description of the industries that the firm works for are connected to associated NAICS codes, using the surrounding context that the terms are used in Wikipedia, to allow terms not directly appearing in the classification manual to be connected. This approach also provides the basis to increase the precision that industries can be identified (e.g., NAICS 4-6 digit level).

ENTITY_TYPE	EXAMPLE	%
GENERIC	Company, business, firm, entity, organization, enterprise, entity	65.0
PROVIDER	provider, company that provides, firm that provides, company which provides	6.6
MANUFACTURER	Manufacturer, company that produces, company that manufactures, company engaged in manufacturing	4.7
DEVELOPER	Developer, company that develops, firm which develops, company engaged in developing	2.5
BANK	Bank, banks	2.4
CONSULTANCY	consultancy	1.9
DISTRIBUTOR	distributor, company that distributes	1.3
SUPPLIER	supplier, suppliers, company that supplies	1.0
RETAILER	retailer	0.8
OPERATOR	operator, company that operates	0.5
PRODUCER	producer	0.5
MARKETER	Marketers, company that markets, firm which markets	0.4
DESIGNER	Designer, company that designs	0.3
CHAIN	chain	0.2
SELLER	Seller, company that sells,	0.2
UTILITY	utility, utilities	0.2
CARRIERS	Carriers, carrier	0.1
PUBLISHER	publisher, firm that publishes	0.1
INDUSTRY	industry, market, sector	n/a

Note: Percentages based upon descriptions of firms that a manager has worked for

Table D.8 Summary of Main Entity Types

[73] While the 'multiple' property has yet to be fully implemented, but will in due course.

```
{
   "ORIGINAL":"gold mining company"
   "NAICS_3_DIGIT":212
   "NAICS_3_DESCRIPTION":"Mining (except oil and gas)"
   "ENTITY_TYPE":"GENERIC|COMPANY"
}

{
   "ORIGINAL":"manufacturer of automotive components"
   "NAICS_3_DIGIT":336
   "NAICS_3_DESCRIPTION":"Transportation equipment manufacturing"
   "ENTITY_TYPE":"MANUFACTURER"
}

{
   "ORIGINAL":"retailer of women's clothing"
   "NAICS_3_DIGIT":448
   "NAICS_3_DESCRIPTION":"Clothing and clothing accessories stores"
   "ENTITY_TYPE":"RETAILER"
}
```


Figure D.18. Examples of the **COMPANY-INDUSTRY_AREA** Sub-Ontology

Management Title

```
MANAGEMENT_TITLE:{
   "ORIGINAL": Original text                                          [String]
   "LEVEL": Seniority level of manager                                [MANAGEMENT_LEVEL]
   "FUNCTIONA_AREA": List of functional areas                         [FUNCTIONAL_AREA]
   "INDUSTRY_AREA": List of industrial areas                          [COMPANY-INDUSTRY _AREA]
   "LOCATION": Specific locations that the manager resides over       [LOCATION]
   "REGION":all non-specific regions, based on the REGION
            property of the CHARACTERIZATION ontology                 [CHARCTERIZATION-REGION[74]]
}
USED_IN: POSITION
```

Figure D.19. Specification of the **MANAGEMENT_TITLE** Sub-Ontology

This sub-ontology represents a manager's job-title, separating out the components of the title into the key dimensions of variation: i) level (e.g., '**Vice President**'), ii) functional area (e.g., '**Marketing**'), iii) industrial area ('**pharmaceutical**') and iv) region (e.g. '**China**').


```
{
   "ORIGINAL":"Vice President of Marketing for the Chinese pharmaceutical division"
   "LEVEL":{"ORIGINAL":"Vice President","LEVEL":"PRESIDENT","Vice":True,"CLEANED":"VICE PRESIDENT"},
   "FUNCTIONAL_AREA":[{"ORIGINAL":"Marketing","PRIMARY_AREA":"MARKETING"}],
   "INDUSTRY_AREA":{"AREA":"PHARMACUTICAL","NAICS_CODE":325},
   "LOCATION":{"COUNTRY":"CHINA"},
}

{
   "ORIGINAL":"Executive Vice President, Corporate Relations"
   "LEVEL":{"ORIGINAL":"Executive Vice President","LEVEL":"PRESIDENT","Vice":True,"EXECUTIVE":":True,
            "CLEANED":"EXECUTIVE VICE PRESIDENT"},
   "FUNCTIONAL_AREA":[{"ORIGINAL":"Corporate Relations","PRIMARY_AREA":"RELATIONS","SUB_AREA":"Corporate"}],
}

{
   "ORIGINAL":"Interim Chief Technology Officer"
   "LEVEL":{"ORIGINAL":"Interim Chief Technology Officer","LEVEL":"CHIEF-OFFICER","INTERIM":True,
            "CLEANED":"CHIEF-OFFICER"},
   "FUNCTIONAL_AREA":[{"ORIGINAL":"Technlogy","PRIMARY_AREA":"TECHNOLOGY"}],
}

{
   "ORIGINAL":"CO-CEO"
   "LEVEL":{"ORIGINAL":"CO-CEO","LEVEL":"CEO","CO-LEVEL":True,"CLEANED":"CEO"},
}
```


Figure D.20. Examples of the **MANAGEMENT_TITLE** Sub-Ontology

[74] Specifically, the levels and classifications are parallel the **REGION** property in the characterizations sub-ontology used to describe firms; other dimensions such as **DIVERSIFICATION** or **SIZE** however are not relevant in the context of managerial titles.

Committee

```
COMMITTEE:{
   "ORIGINAL": Original text                        [String]
   "COMMITTEES": List of committees                 [String - see Table D.9 below]
   "SUB-COMMITTEE": If noted as a sub-committee     [True/False]
}

USED_IN: POSITION
```

Figure D.21. Specification of the COMMITTEE Sub-Ontology

This sub-ontology captures the range of the board's committees that managers can be members of. As illustrated in Table D.9, and Figure D.22, slight variations in committee names are standardized together.

Committee type	Example	%
AUDIT	Audit Committee	32.5
COMPENSATION	Compensation Committee, Remunerations Committee	17.5
NOMINATING	Nominating Committee	12.5
GOVERNANCE	Governance Committee, Corporate Governance Committee	12.4
EXECUTIVE	Executive Committee, Executive Management Committee	7.4
FINANCE	Finance Committee	2.6
INVESTMENTS	Investment committee	2.5
RISK	Risk Committee, Risk Management Committee	1.2
HUMAN_RESOURCES	Human Resources Committee, Personnel Committee, HR Committee	1.0
LOAN	Loan Committee	0.9
ENVIRONMENTAL	Environmental Committee	0.3
OTHERS	AFFAIRS, COMPLIANCE, ETHICS, HEALTH_&_SAFTEY, LIABILITY, OVERSIGHT, PLANNING, POLICY, STRATEGY, TECHNOLOGY	9.3

Table D.9. Primary Committee Types

```
{
   "ORIGINAL":"Compensation, Nominating and Investment Committees",
   "COMMITTEES":["COMPENSATION","NOMINATIONS","INVESTMENT"]
}

{
   "ORIGINAL":"Nominating and Corporate Governance Committee",
   "COMMITTEES":["NOMINATIONS","GOVERNANCE"]
}
```

```
{
   "ORIGINAL":"Investment Committee and the Finance Committee",
   "COMMITTEES":["INVESTMENT","FINANCE"]
}
```


Figure D.22. Examples of the **COMMITTEE** Sub-Ontology

Board

```
BOARD:{
   "ORIGINAL": Original text                                          [String]
   "TYPE": DIRECTORS, GOVERNORS, TRUSTEES, OTHER/NOT_SPECIFIED[75]    [String]
}
USED_IN: POSITION
```

Figure D.23. Specification of the BOARD Sub-Ontology

This ontology is designed to represent the board that a manager serves on.[76] There are inherently very few dimensions to this piece of information, with three types: in the case of companies, boards of directors (DIRECTORS), in the case of universities, boards of governors (GOVERNORS), and in the case of non-profits boards of trustees (TRUSTEES).

```
{
   "ORIGINAL":"Board of Directors"
   "TYPE":"DIRECTORS"
}

{
   "ORIGINAL":"Board of Governors"
   "TYPE":"GOVENORS"
}

{
   "ORIGINAL":"Board of Trustees"
   "TYPE":"TRUSTEES"
}

{
   "ORIGINAL":"Board"
   "TYPE":"OTHER/NOT_SPECIFIED"
}
```

Figure D.24. Examples of the BOARD Sub-Ontology

[75] While there are instances in which the board type is not specified, since in the majority of board types are over companies, it is reasonable to assume that at least in the majority of cases the board type is also DIRECTORS. In due course, it is intended that there will be an option to populate the type based on the organizational form that the board presides over.

[76] It should be noted that while boards are clearly an important aspect of the positions that a manager can serve, and boards are distinct from other information types (e.g., COMMITTEES), unlike most concepts there is inherently little variation in board types. While the data has been split out by type for completeness (e.g., DIRECTORS, GOVERNORS and TRUSTEES), the vast majority of occurrences are DIRECTORS and the difference between the board types at least in part a reflecting merely a different label depending on the organizational form that the board is presiding over (although see: Fama and Jensen, 1983, which although indicating that boards play similar oversight roles across companies, non-profits and universities nevertheless also notes differences in the nature of the oversight:).

Firm Financial

```
{
   "ORIGINAL": Original text                                              [String]
   "CURRENCY": DOLLAR, EURO                                               [String]
   "UNITS_MILLION": Number of units, converted to millions                [Float]
   "OVER": Over/more than                                                 [True/False]
   "AROUND": Around/approximately                                         [True/False]
   "VALUE_TYPE": VALUE, REVENUE, ASSETS                                   [String]
}
USED_IN: COMPANY_DESCRIPTION
```

Figure D.25. Specification of the **FIRM_FINANCIAL** Sub-Ontology

This sub-ontology is intended to represent any mention of the financial position of organizations that a manager has worked for, converting numbers into a standardized format. It should be noted, that in the vast majority of cases, the values are specified without clearly specifying what the value corresponds to, e.g., '**a $1.5 billion dollar retailer**'. While this value may well correspond to the market value of the firm, there is inherent ambiguity, and maybe especially for private firms, where the value is less objectively ascertained. This sub-ontology reflects the inherent ambiguity in the text.

```
{
   "ORIGINAL":"$11 billion",      (e.g., a $11 billion company)
   "CURRENTY":"DOLLAR",
   "UNITS_MILLION":11000
}

{
   "ORIGINAL":"approximately $330 million",
   "CURRENTY":"DOLLAR",
   "AROUND":True
   "UNITS_MILLION":330
}

{
   "ORIGINAL":"revenue of $200 million",
   "CURRENTY":"DOLLAR",
   "AROUND":True
   "VALUE_TYPE":"REVENUE",
   "UNITS_MILLION":200
}
```

Figure D.26. Examples of the **FIRM_FINANCIAL** Sub-Ontology

Listing-Ownership

```
LISTING-OWNERSHIP:{
   "ORIGINAL": Original text                                    [String]
   "OWNERSHIP_TYPE":"PUBLICALLY_LISTED"/"PRIVATE"/"STATE"       [String]
   "EXCHANGE":[{
        "EXCHANGE_NAME":                                        [String - see examples in Table D.10]
        "COUNTRY":                                              [String]
        ]}}
USED_IN: COMPANY_DESCRIPTION
```

Figure D.27. Specification of the LISTING-OWNERSHIP Sub-Ontology

This sub-ontology is designed to capture the listing and ownership details of a company, typically used when describing companies that a manager has worked at, capturing and standardizing information on the exchanges that the company is listed on. If an exchange is provided, then the firm is considered publically listed. As indicated in Table D.10, sub-markets on an exchange (e.g., the 'Alternative Investment Market', a sub-market on the London Stock Exchange, or the 'Toronto Venture Exchange', a sub-market on the Toronto Stock Exchange) are given standardized names that make their connections to the primary exchange clear. The country is populated by looking up the exchange.

EXCHANGE_NAME	COUNTRY	EXAMPLES
NYSE	USA	New York Stock Exchange, NYSE
NASDAQ	USA	NASDAQ, NASDAQ_Stock exchange, NASDAQ_Stock Market
LSE	UNITED_KINGDOM	London_Stock_Exchange
LSE|AIM	UNITED_KINGDOM	AIM, Alternative Investment Market
TSX	CANADA	Toronto Stock Exchange, Toronto Exchange
TSX|VENTURE	CANADA	Toronto Venture Exchange

Note: At present, details are included for approximately 40 exchanges which encompass essentially the entirety of the exchanges discussed in managerial backgrounds since 2007 (with North American exchanges naturally being by far the most commonly discussed). In due course, it is anticipated that this will be expanded to encompass the spectrum of exchanges globally.

Table D.10. Summary of Stock Exchanges

```
{
   "ORIGINAL":"listed on the New York Stock Exchange",
   "OWNERSHIP_TYPE":"PUBLICALLY_LISTED",
   "EXCHANGE":[{"EXCHANGE_NAME":"NYSE","COUNTRY":"USA"}]
}
```

```
{
   "ORIGINAL":"traded on both the NASDAQ and the London Stock Exchange",
   "OWNERSHIP_TYPE":"PUBLICALLY_LISTED",
   "EXCHANGE":[{"EXCHANGE_NAME":"NASDAQ","COUNTRY":"USA"},
               {"EXCHANGE_NAME":"LSE","COUNTRY":"UNITED_KINGDOM"}]
}
{
   "ORIGINAL":"a public firm",
   "OWNERSHIP_TYPE":"PUBLICALLY_LISTED"
}
{
   "ORIGINAL":"a state-owned company",
   "OWNERSHIP_TYPE":"STATE"
}
```


Figure D.28. Examples of the `LISTING-OWNERSHIP` Sub-Ontology

Characterizations

```
{
   "ORIGINAL": Original text                          [String]
   "SIZE": Original text                              [String, see Table D.11]
   "REGION": Original text                            [String, see Table D.11]
   "LEADING": Original text                           [String, see Table D.11]
   "DIVERSIFICATION": Original text                   [String, see Table D.11]
   "PROFIT_STATUS": Original text                     [String, see Table D.11]
   "STAGE": Original text                             [String, see Table D.11]
}
USED_IN: COMPANY_DESCRIPTION, EXPERIENCE, MANAGEMENT_TITLE
```

Figure D.29. Specification of the CHARACTERIZATION Sub-Ontology

This sub-ontology is designed to represent qualitative characterizations, which, within the context of managerial backgrounds, are used to describe the organizations that the manager has worked for, and the experience that a manager has gained. The intention is that this ontology will provide the basis to represent a broader range of qualitative descriptions that may, for example, be used to describe products, or used by stakeholders to describe firms (which may include negative descriptions). As such, while the dimensions characterized in this ontology reflect the dimensions that are discussed within managerial backgrounds, more dimensions will be added in due course to increase the flexibility of this ontology to capture a broader spectrum of discussion.

PROPERTY	VALUE	Examples	% OF PROPERTY
SIZE	LARGE	largest, major, large, one of the largest	77.9
	MEDIUM	medium, mid-sized	0.9
	SMALL	small, smallest	21.2
REGION	GLOBAL/INTERNATIONAL	global, international, worldwide, multinational	73.5
	NATIONAL	national, nationwide	19.6
	REGIONAL/LOCAL	local, regional	6.9
LEADING	LEADING	leading, premier, world-leading, one of the leading	n/a
DIVERSIFICATION-FOCUS	DIVERSIFIED	diversified	100
	FOCUS	focused	negligible
PROFIT_STATUS	NOT_FOR_PROFIT	non-profit, charitable, not-for-profit	99.8
	FOR_PROFIT	for profit, for-profit	0.2[77]

[77] For-profit organizations are rarely explicitly labeled as such (while non-profits are more commonly labeled non-profits); this explains why the non-profit label more commonly occurs in describing the organizations worked for than the for-profit.

STAGE	EARLY/START_UP	start-up, early-stage	45.2
	GROWING	Growing	21.3
	BOUTIQUE	boutique	12.8
	EMERGING	Emerging	10.5
	DEVELOPMENT	Development stage	3.3
	OTHER	pre-IPO, clinical-stage	6.8

Table D.11. Summary of Characterization Areas


```
{
   "ORIGINAL":"leading, diversified"
   "LEADING":"LEADING"
   "DIVERSIFICATION-FOCUS":"DIVERSIFIED"
}

{
   "ORIGINAL":"one of the largest"
   "SIZE":"LARGE"
}
```


Figure D.30. Examples of the CHARACTERIZATION Sub-Ontology

Company Description

```
COMPANY_DESCRIPTION:{
  "ORIGINAL": Original text                                              [String]
  "INDUSTRY": List of industrial areas                                   [LIST: INDUSTRY]
  "REGION": Regions that the listed as operating in                      [LIST: LOCATION]
  "FIRM_FINANCIAL": Description of the value of firm                     [FIRM_FINANCIAL]
  "LISTING-OWNERSHIP": Listing/exchange, and ownership details           [LISTING-OWNERSHIP]
  "CHARACTERIZATION": Qualitative discussion of the firm (e.g., leading) [CHARACERIZATION]
  "FOUNDED_DATE": A date that the firm is listed as founded              [DATE]
}
USED_IN: POSITION
```

Figure D.31. Specification of the COMPANY_DESCRIPTION Sub-Ontology

This sub-ontology represents the qualitative description of the firms that a manager has worked for. This description is broken down by details of the INDUSTRY in which the firm is discussed as operating in, any REGION that is discussed, the STAGE of the firm, and other CHARACTERIZATION, where other subjective characteristics such as whether the firm is "leading" are discussed.

```
{
   "ORIGINAL":"NYSE-listed gold mining company"
   "INDUSTRY":{"NAICS_3_DIGIT":212,"NAICS_3_DESCRIPTION":"Mining (except oil and gas)"}
   "LISTING-OWNERSHIP:{"OWNERSHIP_TYPE":"PUBLICALLY_LISTED"
                       "EXCHANGE":[{"EXCHANGE_NAME":"NYSE","COUNTRY":"USA"}]
}

{
   "ORIGINAL":"leading US retailer of women's clothing"
   "INDUSTRY":{"NAICS_3_DIGIT":212, "NAICS_3_DESCRIPTION":"Mining (except oil and gas)"}},
   "REGION":[{"COUNTRY":"USA"}]
   "CHARACTERIZATION":[{"TERM":"leading","AREA":"LEADING"}]
}
```

Figure D.32. Examples of the COMPANY_DESCRIPTION Sub-Ontology

Company Name

```
COMPANY_NAME:{
   "ORIGINAL": Original text                                          [String]
   "CLEANED": Company name with endings (e.g., Inc., Corp.,) cleaned  [String]
}
USED_IN: POSITION
```

Figure D.33. Specification of the COMPANY_NAME Sub-Ontology

This sub-ontology is designed to represent company names. As well as the original company name, a cleaned version is included, removing organization endings (e.g., Inc., Corp, etc.) to facilitate connections between slight permutations in the way organizational names are written.[78] As illustrated below, if the focal company is referred to, then the cleaned version will be "FOCAL_COMPANY".

```
{
   "ORIGINAL":"Gorilla Software Inc."
   "CLEANED":"GORILLA SOFTWARE"
}

{
   "ORIGINAL":"Oculi Machined Parts Corp"
   "CLEANED":"OCULI MACHINED PARTS"
}

{
   "ORIGINAL":"our company"
   "CLEANED":"FOCAL_COMPANY"
}
```

Figure D.34. Examples of the COMPANY_NAME Sub-Ontology

[78] In due course it is intended to connect the company names to other databases, making it easier to integrate the extracted information to external databases.

Person Name

```
PERSON_NAME:{
   "ORIGINAL": Original text                                    [String]
   "NAME_TITLE": Name title, standardized to MR, MS, DR, MRS    [String]
   "FIRST_NAMES": First name (including initials)               [String]
   "LAST_NAME": Last name                                       [String]
   "SUFFIX": Suffix                                             [String]
}
USED_IN: BACKGROUND_DETAILS
```

Figure D.35. Specification of the PERSON_NAME Sub-Ontology

This sub-ontology is designed to represent people names, splitting the name up by part, such that it is possible to connect individuals with the same name, irrespective of whether the name is written in exactly the same format in subsequent filings. The name is split into appropriate properties based on the sequencing of the name, following a standard structure. Some of the most common sequences being: i**)** NAME_TITLE FIRST_NAMES LAST_NAME (e.g., 'Mr. John Doe'**)**, ii) FIRST_NAME LAST_NAME, (e.g., 'John Doe') iii), NAME_TITLE LAST_NAME, (e.g., 'Mr. Doe') iv) FIRST_NAME) (e.g., 'John'). For example, if the name contains a name title followed by several words (e.g., 'Mr. John Doe'), then 'Mr.' is taken as the NAME_TITLE, the next word (John) is taken as the FIRST_NAME, and the last word ('Doe') is taken as the LAST_NAME. Similarly, if just one word is given (e.g., 'John'), this is taken as the first name.[79]

```
{
   "ORIGINAL":"Mr. John F. Doe III"
   "NAME_TITLE":MR"
   "FIRST_NAMES":"JOHN F."
   "LAST_NAME":"DOE"
   "SUFFIX":"III"
}
{
   "ORIGINAL":"Mrs. Doe"
   "NAME_TITLE":MRS"
   "LAST_NAME":"DOE"
}
{
   "ORIGINAL":"Jane Doe"
   "FIRST_NAMES":"JANE"
   "LAST_NAME":"DOE"
}
{
   "ORIGINAL":"John"
   "FIRST_NAMES":"JOHN"
}
```

Figure D.36. Examples of the PERSON_NAME Sub-Ontology

[79] It is common to specify a manager's full name in the first instance in the managerial background, and then either use their first name or 'he'/'she'. That is, while only knowing the first name may not be sufficient to uniquely identify a manager, the full name is typically mentioned several sentences earlier.

Background Details

```
BACKGROUND_DETAILS:{
   "ORIGINAL_TEXT": Original text                                                    [String]
   "PERSON_NAME": Name split up                                                      [PERSON_NAME]
   "AGE": Age in years                                                               [Integer]
   "POST_NOMINAL_LETTERS": All nominal letters appearing after a name
   (i.e., qualification "letters", that are not the focus of the sentence)  [see example below]
}
```

Figure D.37. Specification of the BACKGROUND_DETAILS Primary-Ontology

This structure is designed to capture descriptive information about the manager not typically the focus of the sentence. This includes post-nominal letters that are sometimes included after a manager's name, but are not the focus of the sentence, which are standardized in this ontology in a similar manner to how qualifications/professional license are captured. [80]

```
BACKGROUND_DETAILS:{
   "ORIGINAL_TEXT":"Mr. John Doe",
   "PERSON_NAME":{"ORIGINAL_TEXT":"Mr. John Doe","TITLE":"MR","FIRST_NAME":"JOHN","LAST_NAME":"DOE"}
}

BACKGROUND_DETAILS:{
   "ORIGINAL_TEXT":"Mrs. Jane Doe, age 56",
   "PERSON_NAME":{"ORIGINAL_TEXT":"Mr. Jane Doe","TITLE":"MRS","FIRST_NAME":"JANE","LAST_NAME":"DOE"}
   "AGE":{"ORIGINAL_TEXT":"age 56" "AGE":56}
}

BACKGROUND_DETAILS:{
   "ORIGINAL_TEXT":"Mr. John Doe, PhD, (age 64)",
   "PERSON_NAME":{"ORIGINAL_TEXT":"Mr. John Doe","TITLE":"MR","FIRST_NAME":"JOHN","LAST_NAME":"DOE"}
   "POST_NOMINAL_LETTERS":[{"ORIGINAL_TEXT":"PhD","TYPE":"QUALIFICATION",
                            "LEVEL":"POST_GRADUATE/DOCTORATE"}]
   "AGE":{"ORIGINAL_TEXT":"(age 64)" "AGE":64 }
}
```

Figure D.38. Examples of the BACKGROUND_DETAILS Primary-Ontology

[80] These qualifications also tend to be discussed separately as part of a later sentence.

Position

```
POSITION:{
   "ORIGINAL_TEXT": Original text                                  [String]
   "JOB_TITLE":                                                    [MANAGEMENT_TITLE]
   "COMPANY_NAME":                                                 [COMPANY_NAME]
   "COMPANY_DESCRIPTION":                                          [COMPANY_DESCRIPTION]
   "BOARD":Incorporates a manager's role on the board              [BOARD]
   "COMMITTEE":Incoporates a manager's role on the committee       [COMMITTEE]
   "START_DATE":                                                   [DATE]
   "END_DATE":                                                     [DATE]
}
USED_IN: POSITIONS
```

Figure D.39. Specification of the POSITION Sub-Ontology

This data structure integrates the components of an individual position, including the committees and boards that a manager has worked on (described below).

```
POSITION:{
   "ORIGINAL":"Chief Financial Officer of Gorilla Software Inc. from April 1997 until March 2000",
   "JOB_TITLE":{"ORIGINAL":"Chief Financial Officer","LEVEL":"CHIEF-OFFICER","AREA":["FINANCE"]}
   "COMPANY_NAME":{"ORIGINAL":"Gorilla Software","NAME_CLEANED":"GORILLA SOFTWARE"}
   "START_DATE":{"ORIGINAL":"April 1997","YEAR":1997,"MONTH":4}
   "END_DATE":{"ORIGINAL":"March 2000","YEAR":2000,"MONTH":3}
}

POSITION:{
   "ORIGINAL":"between 1997 and 2000 he was the Vice President of OPERATIONS at GORILLA Software Inc.",
   "JOB_TITLE":{"ORIGINAL":"Vice President of OPERATIONS","LEVEL":"CHIEF-OFFICER",
                "AREA":["OPERATIONS"]}
   "COMPANY_NAME":{"ORIGINAL":"Gorilla Software","NAME_CLEANED":"GORILLA SOFTWARE"}
   "START_DATE":{"ORIGINAL":"1997","YEAR":1997}
   "END_DATE":{"ORIGINAL":"2000","YEAR":2000}
}

POSITION:{
   "ORIGINAL":"VP of Marketing and Sales at Gorilla Software Inc. until he retired in 2010",
   "JOB_TITLE":{"ORIGINAL":"VP of Marketing and Sales","LEVEL":"VP","AREA":["MARKETING","SALES"]}
   "COMPANY_NAME":{"ORIGINAL":"Gorilla Software","NAME_CLEANED":"GORILLA SOFTWARE"}
   "END_DATE":{"ORIGINAL":"retired in 2010","YEAR":2000, "EVENT":"RETIREMENT"}
}
```

Figure D.40. Examples of the POSITION Sub-Ontology

For board and committee positions, the position title (e.g., '**member**' or '**chair**') are incorporated into the **BOARD** and **COMMITTEE** sub-ontologies (i.e., because such positions are relative to the committee/board, rather the overall company).[81]


```
POSITION:{
   "ORIGINAL":"member of our Audit Committee",
   "COMPANY_NAME":{"ORIGINAL":"our","NAME_CLEANED":"FOCAL_FIRM"},
   "COMMITTEE":{"ORIGINAL":"member … Audit Committee","COMMITTEES":["AUDIT"],"POSITION":"MEMBER"}

}

POSITION:{
   "ORIGINAL":"chair of our board of directors",
   "COMPANY_NAME":{"ORIGINAL":"our","NAME_CLEANED":"FOCAL_FIRM"},
   "BOARD":{"ORIGINAL":"chair … board of directors","TYPE":"DIRECTORS","POSITION":"CHAIR"}

}
```


Figure D.41. Examples of the **POSITION** Sub-Ontology with Boards/Committees

As illustrated in Figure D.46 below, the **POSITIONS** primary-ontology, is then made up as a list of the **POSITION** type, allowing multiple positions discussed in the same sentence to be represented.


```
{
 "ORIGINAL_SENTENCE":"Mr. Doe previously worked for Gorilla Software Inc. as the Vice President of
                      OPERATIONS between 2001 and 2005, and worked for Oculi Machined Parts as the Head of
                      the Operations Division between 1996 and 2000.",
"BACKGROUND":{"PERSON_NAME":{"ORIGINAL":"Mr. Doe","NAME_TITLE":"MR","LAST_NAME":"DOE"}
"POSITIONS":[
 {
  "ORIGINAL":"worked for Gorilla Software Inc. as the Vice President of OPERATIONS between 2001 and 2005",
  "JOB_TITLE":{"ORIGINAL":"Vice President of OPERATIONS","LEVEL":"VP","AREA":["OPERATIONS"]}
  "COMPANY_NAME":{"ORIGINAL":"Gorilla Software Inc.","NAME_CLEANED":"GORILLA SOFTWARE"}
  "START_DATE":{"ORIGINAL":"2001","YEAR":2001}
  "END_DATE":{"ORIGINAL":"2005","YEAR":2005}
 },
 {
  "ORIGINAL":"worked for Oculi Machined Parts as the Head of the Operations Division between 1996 and 000",
  "JOB_TITLE":{"ORIGINAL":"Head of Marketing","LEVEL":"HEAD","AREA":["MARKETING"]}
  "COMPANY_NAME":{"ORIGINAL":"Oculi Machined Parts","ORIGINAL":"OCULI MACHINED PARTS"}
  "START_DATE":{"ORIGINAL":"1996","YEAR":1996}
  "END_DATE":{"ORIGINAL":"2000","YEAR":2000}
 }]
}
```


Figure D.42. Examples of the **POSITIONS** Primary-Ontology

[81] As of August 1st 2018, this has yet to be implemented, but will in due course (i.e., they are currently treated as any other position).

Experience

```
EXPERIENCE:{
   "ORIGINAL": Original text                                                    [String]
   "LENGTH_OF_TIME":all specific lengths of time e.g., 5 years                  [LENGTH_OF_TIME]
   "EXPERIENCE_TYPE":                                                           [see Table D.12 below]
   "CHARACTERIZATION":all qualitative description of experience                 [CHARACTERIZATION]
   "FUNCTIONAL_AREA: functional areas corresponding to functional background    [FUNCTIONAL_AREA]
   "INDUSTRY":all industries                                                    [INDUSTRY]
   "REGION": all non-specific regions (e.g., 'global'), based on the REGION
             property of the CHARACTERIZATION ontology                          [CHARCTERIZATION-REGION]
   "GEOGRAPHIC_AREA":specific countries / regions mentioned                     [LOCATION]
}
USED_IN:EXPERIENCES
```

Figure D.43. Specification of the **EXPERIENCE** Sub-Ontology

Experience Type	Example usage	%
EXPERIENCE	significant experience in marketing, accounting and economics	75.7
KNOWLEDGE	knowledge of accounting	11.1
BACKGROUND	accounting background	5.0
INSIGHT	broad understanding of the operational, financial and strategic issues	2.4
PERSPECTIVE	unique perspective	2.3
CAREER	career in the engineering, procurement, and construction industry	1.3
QUALIFICATIONS	qualifications [and experience] in accounting	1.3
FAMILIARITY	familiar with the Company's business and industry	0.5
ABILITY	ability to lead	0.1
SKILL	management skills	0.1

Table D.12. Summary of Experience Types

```
EXPERIENCE {"ORIGINAL":"worked in the automotive industry for 5 years",
   "LENGTH_OF_TIME":{"ORIGINAL":"5 years","UNIT":"YEAR","QUANTITY":5},
   "AREAS":{"ORIGINAL":"automotive industry","NAICS_3_DIGIT":336,
           "NAICS_3_DESCRIPTION":"Transportation Equipment Manufacturing"},
           "EXPERIENCE_TYPE":{"ORIGINAL":"worked","TYPE":"WORK"}
}

EXPERIENCE:{
   "ORIGINAL":"20 year career in the global oil and gas industry",
   "LENGTH_OF_TIME":{"ORIGINAL":"5 years","UNIT":"YEAR","QUANTITY":5},
   "AREAS":{"ORIGINAL":"oil and gas industry","NAICS_3_DIGIT":211,
            "NAICS_3_DESCRIPTION":"Oil and Gas Extraction"},
   "REGION":{"ORIGIINAL":"global","TYPE": GLOBAL/INTERNATIONAL"},
   "EXPERIENCE_TYPE":{"ORIGINAL":"career","TYPE":"CAREER"}}
```

Figure D.44. Examples of the **EXPERIENCE** Sub-Ontology

Qualification

```
QUALIFICATION:{
   "ORIGINAL": Original text                    [String]
   "DEGREE": Original text                      [DEGREE]
   "EDUCATION_INSTITUTION": Original text       [EDUCATION_INSTITUTION]
   "GRADUATION_DATE": Original text             [DATE]
}
USED_IN: QUALIFICATIONS
```

Figure D.45. Specification of the QUALIFICATION Sub-Ontology

This sub-ontology represents a single qualification, made up of the components the DEGREE, EDUCATION_INSTITUTION, and DATE.

```
{
   "ORIGINAL":"graduated from the University of Michigan in 1987 with a BBA"
   "DEGREE":{"ORIGINAL":"BBA","DEGREE_LEVEL":"UNDERGRADUATE/BACHELORS","SUBJECTS":["BUSINESS"]}
   "EDUCATION_INSTITUTION":{"ORIGINAL":"University of Michigan","UNIVERSITY":"UNIVERSITY OF MICHIGAN"}
   "GRADUATION_DATE":{"ORIGINAL":"1987","YEAR": 1987}
}

{
   "ORIGINAL":"received a BEng in 1990 with a from the University of Illinois"
   "DEGREE":{"ORIGINAL":"BEng","DEGREE_LEVEL":"UNDERGRADUATE/BACHELORS","SUBJECTS":["ENGINEERING"]}
   "EDUCATION_INSTITUTION":{"ORIGINAL": University of Illinois","UNIVERSITY":"UNIVERSITY OF ILLINOIS"}
   "GRADUATION_DATE":{"ORIGINAL":"1990","YEAR": 1990}
}

{
   "ORIGINAL":"earned a Bachelor of Business Administration from the University of Stanford in 1996"
   "DEGREE":{"ORIGINAL":"Bachelor of Business Administration",
             "DEGREE_LEVEL":"UNDERGRADUATE/BACHELORS","SUBJECTS":["BUSINESS"]}
   "EDUCATION_INSTITUTION":{"ORIGINAL":"University of Stanford","UNIVERSITY":"STANFORD UNIVERSITY"}
   "GRADUATION_DATE":{"ORIGINAL":"1996","YEAR": 1996}
}
```

Figure D.46. Examples of the QUALIFICATION Sub-Ontology

As illustrated in Figure D.47 below, the QUALIFICATIONS primary-ontology, is then made up as a list of the QUALIFICATION type, allowing multiple qualifications discussed in the same sentence to be represented.

```
{
  "ORIGINAL_SENTENCE":"Mr. Doe earned a Bachelor of Science in engineering from the University of
                       Michigan in 1980 and a Master of Business Administration from the Stanford Graduate
                       School of Business in 1985.",
  "QUALIFICATIONS":[
      {
        "EDUCATION_INSTITUTION":{"ORIGINAL":"University of Michigan","UNIVERSITY":"UNIVERSITY OF MICHIGAN"},
        "DEGREE":{"ORIGINAL":"Bachelor of Science in engineering",
                  "LEVEL":"UNDERGRADUATE/BACHELORS","SUBJECTS":["ENGINEERING"]}
        "GRADUATION_DATE":{"ORIGINAL":"1980","YEAR": 1980}
      },
      {
        "EDUCATION_INSTITUTION":{"ORIGINAL":"Stanford Graduate School of Business"
                           "UNIVERSITY":"STANFORD UNIVERSITY","DEPARTMENT":"Graduate School of Business"}
        "DEGREE":{"ORIGINAL":"Master of Business Administration"
                  "LEVEL":"GRADUATE/MASTERS","SUBJECTS":["BUSINESS"]}
        "GRADUATION_DATE":{"ORIGINAL":"1985","YEAR": 1985}
      }
   ]
}
```


Figure D.47. Examples of the **QUALIFICATIONS** Primary-Ontology

Professional License

```
PROFESSIONAL_LICENSE:{
   "ORIGINAL": Original text                                    [String]
   "AREA": Area that the license is in                          [String - see Table D.13 below]
   "REGIONS": List of all regions that they are licensed in     [LOCATION[82]]
   "INACTIVE": e.g., inactive/no longer active                  [True/False]
}
USED_IN: PROFESSIONAL_LICENSES
```

Figure D.48. Specification of the **PROFESSIONAL_LICENSE** Sub-Ontology

This sub-ontology is designed to represent professional licenses that a manager may hold. Depending on the profession, such designated included 'licensed' (e.g., '**licensed to practice law**'), 'registered' (e.g., '**registered professional engineer**'), '**certified**' (e.g., '**certified public accountant**') or '**chartered**' (e.g., '**chartered surveyor**'), with the terms often stating a particular state or country in which they are valid. Although these could be considered a form of education, they are treated separately, because unlike other qualifications they are not issued by universities (and typically the entity issuing the license is not directly mentioned), often mention the region in which they are valid, and typically confer some legal responsibility (e.g., the ability to certify accounts).

[82] For simplicity, the **LAT-LONG** and the **ORIGINAL** properties of the **REGION** sub-ontology have been are omitted here; the intention is that ultimately the tools will have options to allow uncommon/undesired properties to be suppressed/unsuppressed.

Area	Examples
ACCOUNTING|CPA[83]	CPA, CPA certificate
ACCOUNTING|CMA[83]	CMA, Certified Management Accountant
FINANCE|CFA[83]	Licensed Chartered Financial Analyst, CFA Charter Holder
ENGINEER	Licensed Professional Engineer, Registered Professional Engineer
LAWYER	Licensed to Practice Law
PHARMACIST	Registered Pharmacist
ARCHITECT	Registered Architect
NURSE	Registered Nurse
REAL_ESTATE_BROKER	Registered Real Estate Broker
PHYSICIAN	Registered Pharmacist
SURVEYOR	Chartered Surveyor

Table D.13. Summary of Professional License Areas


```
{
   "ORIGINAL":"licensed to practice law in New York and New Jersey",
   "AREA":"LAW",
   "REGION_LIST":[{"COUNTRY":"USA","STATE":"NY"},{"COUNTRY":"USA","STATE":"NJ"}]
}

{
   "ORIGINAL":"inactive C.P.A License in Missouri",
   "AREA":"ACCOUNTING|CPA",
   "REGION_LIST":[{"COUNTRY":"USA","STATE":"MO"}]
   "INACTIVE": True
}

{
   "ORIGINAL":"licensed as a Chartered Financial Accountant in the state of Michigan",
   "AREA":"ACCOUNTING|CFA",
   "REGION_LIST":[{"COUNTRY":"USA","STATE":"MI"}]
}

{
   "ORIGINAL":"Registered nurse in Canada and the United Kingdom",
   "AREA":"NURSE",
   "REGION_LIST":[{"COUNTRY":"CANADA"},{"COUNTRY":"UNITED_KINGDOM"}]
}
```


Figure D.49. Examples of the PROFESSIONAL_LICENSE Sub-Ontology

[83] There are two common accounting certificates, the CPA (Certified Public Accountant) and CMA (Certified Management Accountant), as well as the more finance-focused CFA (Chartered Financial Analyst). While each is split out separately here due to differences in content between the qualifications (The Constant Analyst, 2013), with a label indicating whether they have a finance or accounting focus, the categories can easily be aggregated together as deemed appropriate.

Appendix E

Standardization and Verification

Summary of Standardization and Verification Status

This appendix describes the current status of each of the sub-ontologies, while describing the path to future developments. As noted in the text, many of the concepts are comprised of very large number of terms, with a long tail of relatively uncommon terms. For example, Figure E.1 illustrates the cumulative occurrence of the first 10,000 terms in the **MANAGEMENT_TITLE** concept. As illustrated, while a large proportion of terms are common, a sizable percentage of management titles are comprised of low-frequency terms. Since sentences typically comprise many separate concepts, unless classifications within concepts are very high, (which necessitates capturing the low-frequency terms), mistakes at the sentence level would be common. As described in the text however, the process of dissecting the managerial backgrounds is able to automatically characterize these low frequency terms, and while there is a slight divergence at the low-frequency end of the graph between the fully-dissected terms (blue line) and the overall numbers of terms (brown line), the majority of terms are fully captured.[84]

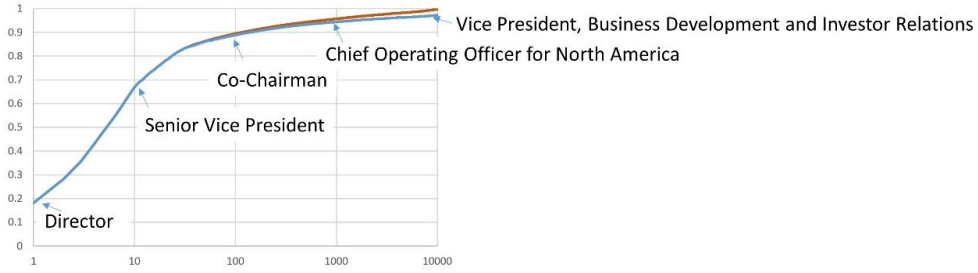


Figure E.1 Cumulative Distribution of Managerial Titles (Logarithmic Scale)

[84] As noted previously, fully dissected terms refers to splitting terms to the underlying properties of the ontologies, (i.e., every word classed to sub-concepts, with the sequencing of those sub-concepts conforming to the expectations of the semantic-interpretation). Effort is on-going to fully dissect remaining terms in the tails; while there are various unusual titles, such as '**Chief of Staff**' or '**Branch Manager**', that do not neatly conform to the existing ontology, it is expected that the rate of classification will ultimately exceed 99%.

As discussed in the main text, while dissecting each of the concepts to the sub-ontologies is an integral to populating the ontologies, it also ensures the validity of the classifications, and the flexibility of the ontologies to capture the underlying dimensions of the text. As illustrated in Table E.1 below, in addition to standardizing the text, fully dissecting terms helps ensures various potential issues involves in the interpretation process are addressed.[85]

Issue	Example	Resolution
Incorrect concept ordering	MANAGEMENT_TITLE: Vice President LEVEL of OF Gorilla Software COMPANY_NAME	Terms with incorrect concept orderings flagged and removed from the concept
Ontology not sufficiently flexible	MT_LVEL: Interim Vice VICE President PRESIDENT	Ontology extended to incorporate dimensions as needed
Semantic rules not sufficiently flexible	DEGREE: Engineering SUBJECT Bachelors LEVEL suma cum laude GRADE_NOTES	Increase flexibility of the semantic-rules to allow a broader number of terms to be represented.
Terms not included in dimensions	DEGREE: Bachelors LEVEL in IN Egyptian Studies UNCLASSIFIED	Increase terms in each dimension, to allow classifications.

Table E.1 Sources of Non-Conformance to Ontologies

The standardization process involves addressing each of these issues, either by removing incorrectly classified terms, increasing the flexibility of the ontologies, or extending the semantic interpretations and classification of terms in each dimension. While a general high level of oversight is maintained across all terms in concepts (with the multiple layers to the validation approach discussed in the text), being able to identify and focus on the proportion of terms most likely to have various forms of mistakes is critical to allowing manual oversight across large volumes of terms.

To illustrate the feasibility of the overall approach to classify a broad array of terms, an initial target was set at 90% of all terms to be fully-classified. As illustrated in Table E.2, this

[85] While conforming to the expected dimensions of the ontologies do not necessarily mean that the information is correctly extracted. For example, if sub-concepts include incorrect terms, it is possible for items to be fully classified, while still having mistakes in how they are populated to the sub-ontologies. While this is possible, and effort is made to ensure that this does not occur, there is little to no evidence of this form of error. More broadly, while this appendix is intended to provide an illustration of the success of the current approach, many additional layers of verification are gradually getting incorporated, with the intent that ultimately it is possible to achieve over 99% of sentences fully and correctly dissected, and substantially higher than would be feasible with manual coding.

target has been satisfied across many concepts, with other concepts expected to reach this target shortly, and in turn, the targets gradually increased.

Sub-ontology	Number of unique terms (2007-2017)	Percentage of occurances classified
DATE	17,374	99.9
LENGTH_OF_TIME	597	95.5
DEGREE	18,356	91.5
EDUCATION_INSTITUTION	4,216	93.8
LOCATION	407	86.8
FUNCTIONAL_AREA	1,942	97.3
MANAGEMENT_LEVEL	1,129	96.2
MANAGEMENT_TITLE	58,122	85.8
COMMITTEE	3,358	87.4
BOARD	299	99.9
COMPANY-INDUSTRY_AREA	25,892	n/a
FIRM_FINANCIAL	163	86.2
COMPANY DESCRIPTION	40033	81.3
LISTING-OWNERSHIP	303	86.5
CHARACTERIZATIONS	594	81.2
PERSON_NAME	141,641	n/a
COMPANY_NAME	137,917	n/a
PROFESSIONAL_LICENSE	1,086	97.9

Table E.2 Proportion and Number of Terms Classified by Sub-Ontology Type

Approaches Under Development to Classify and Verify Fuzzy Concepts

While the approach to dissect terms to properties is appropriate with many concepts, and particularly so for modifiers that can be represented as dichotomous values (e.g., `True/False`), for certain concepts, there is inherent ambiguity over classifications. As discussed in the text, concepts can be ‘fuzzy’ (Murphy, 2002), with less defined boundaries. For certain concepts, pre-existing classification schemas exist that are widely used (e.g., NAICS codes). For other concepts, the coding schemas are less defined; while people may have a conception of categories for university subjects (e.g., sociology, chemistry, psychology), there are different ways of grouping such subjects, and although there will be commonality between universities in departmental groupings of subjects, there will also be variations. The guiding approach intended to serve as the basis for classifications, is to reflect typical groupings; for university subjects, the

typical departments under which the subjects fall. To reflect groupings of university subjects, the envisioned approach will be to download every department website across universities (e.g., top 100 or 500 universities, representing the majority of universities attended by the top managers), and use the department most typically associated with the subject. This builds on the approach discussed to classify NAICS codes below, with the attempt to capture an aggregate of the underlying groupings. For concepts where online-textual data is not amenable to categorize the items, a more traditional classification approach will be used, using crowdsourced verification on a on-line platform, such as Amazon-Turk, to group terms to concepts (e.g., Kittur, Chi, and Suh, 2008), which has scalability advantage over research-assistant coders.

Initial Development of Automated Standardization of Industry and Company Areas to NAICS Codes

As noted in the text, while it was possible to standardize some of the information types through manual consideration, for some information types, there were too many terms for this to be a feasible approach (for example, the tens of thousands of distinct company areas, such as '`distributor of engineering plastics`', '`lighting manufacturing company`', '`men's and women's apparel retailer`', '`developer of automatic data capture software`' etc.). To allow terms to be connected to appropriate NAICS codes, irrespective of whether the term directly appeared, an approach is being developed to automatically connect terms to an appropriate concept, irrespective of whether the specific term itself appeared in the coding manual, using the context surrounding the discussion of the term on a large encyclopedia (namely Wikipedia) to infer the most relevant area.[86] As described in Figure E.2, by using the terms included in the index of the NAICS classification manual, and the surrounding discussion of the company's area on Wikipedia, it is possible to infer a likely NAICS code. In the illustrated example, the phrases '`manufacturer of Application Specific Integrated Circuit`' can be connected to the NAICS code 334, 'Computer and Electronic Product Manufacturing' (with

[86] It should be noted for the outset, that while this approach shows significant promise, and a high proportion of terms are associated with appropriate NAICS codes, the approach is still being refined. While the number of misclassifications is currently higher than desired (e.g., roughly 20% miscategorized), various refinements are yielding improvements. Specifically, using the firm's overall area to restrict classifications to specific NAICS ranges (e.g., restricting firms described as 'manufacturer' to codes 31-33) are gradually improving the classifications; while it is still too early to merit more validation, with further improvements envisioned before this is worthwhile, the more systematic validation is intended in due course.

other permutations such as '**manufacturer of ASICs**') matching up similarly. Thus, while neither of these terms directly appear in the NACIS classification manual (and may be hard for manual coders to connect), they are associated with an appropriate category.

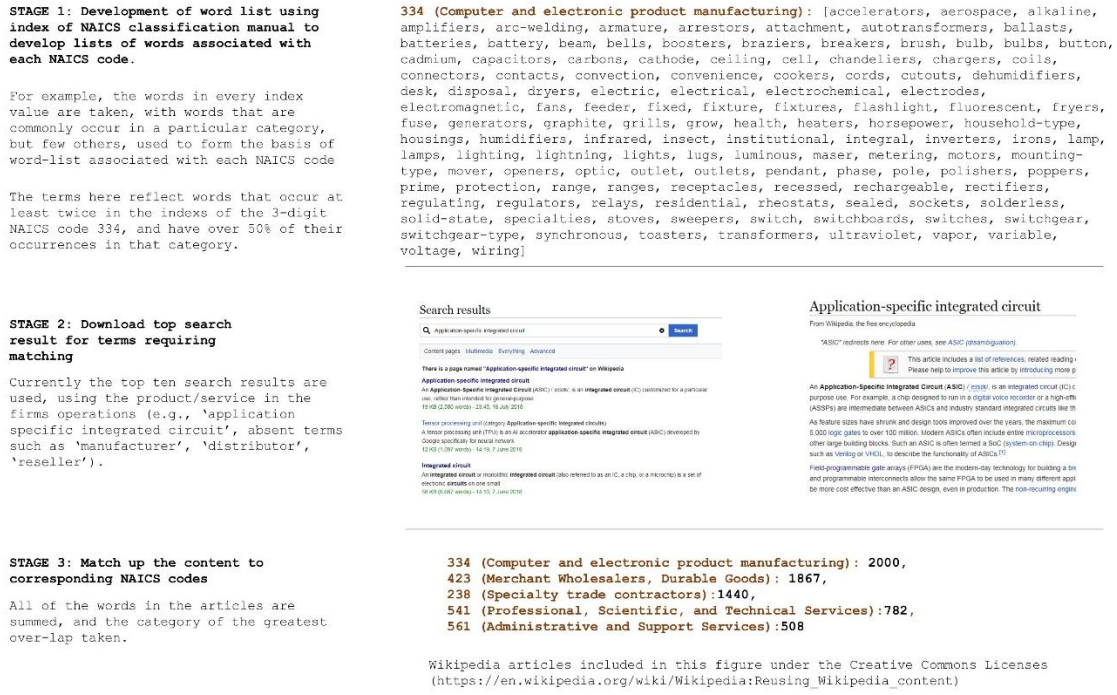


Figure E.2 Illustration of Connecting Company Areas to Associated NAICS codes through Wikipedia Data

Appendix F

Expansion of the Approach

Expansion to Capture Top Manager Selection Decisions

While the ontologies developed in this dissertation are in the context of managerial backgrounds, the overall approach is flexible and can be extended to other situations where there exists some underlying deep-level similarity between the materials. As discussed in the text, not only does expanding to other textual sources allow different forms of questions to be examined, but it also enables more complex theoretical questions, such as how discussion in one medium influences discussion in a different medium. For example, once it is possible to capture representations of discussion of the selection process to appoint top managers, it becomes possible to examine the reciprocal process by which evaluation criteria influence how managerial backgrounds are presented. Figure F.1 below shows a mock-up of a possible extension of the approach to represent discussions of selection procedures. This figure also illustrates the flexibility of the approach to other communication mediums, supplementing the sub-ontologies developed in this dissertation to capture other domain-specific material.

```
{
 "ORIGINAL_SENTENCE":"Candidates may come to the attention of the Nominating Committee through executive
                     search firms, stockholders, management and board members.",
 "COMMITTEE_BACKGROUND":{"COMMITTEE_NAME":{"ORIGNAL":"Nominating Committee","COMMITTEE":["NOMINATING"]}}
 "AWARENESS_CANDIDATES":[{"ORIGINAL":"search firms, stockholders, management and board members"
                          "AWARNESS_FROM":[{"ORIGNAL":"executive search firms","TYPE":"SEARCH_FIRM"},
                                           {"ORIGNAL":"stockholders","TYPE":"SHAREHOLDERS"},
                                           {"ORIGNAL":"management","TYPE":"MANAGMENT"},
                                           {"ORIGNAL":"board members","TYPE":"BOARD"}]
                         }]
}
"
{
 "ORIGINAL_SENTENCE":"When considering a non-incumbent candidate, the Nominating Committee will take into
                     integrity, educational background and knowledge of our business",
"COMMITTEE_BACKGROUND":{"COMMITTEE_NAME":{"ORIGNAL":"Nominating Committee","COMMITTEE":["NOMINATING"]}}
"DECISION_CRITERIA":[{"ORIGINAL":"integrity, educational background and knowledge of our business"}
               "CONSIDERATION_OF_WHO":[{"ORIGNAL":"non-incumbent candidate","TYPE":"NON_INCUMBENT"}],
               "DECISION_CRITERIA":[{"ORIGNAL":"integrity","TYPE":"INTEGRITY"},
                                    {"ORIGNAL":"educational background","TYPE":"EDUCATION"},
                                    {"ORIGNAL":"knowledge of our business","TYPE":"KNOWLEDGE_OF_BUSNESS"},

                         }]
}
```


Figure F.1. Mocked-Up of Ontologies for Selection Decisions

Multi-Lingual Representations

The approach developed in this dissertation is also well suited for capturing representations of sentences in languages other than English. While the specific words differ by language and the way in which sentences are constructed can also vary (for example, adjectives usually come before nouns in English, but afterwards in Spanish: Stockwell, Bowen, and Martin, 1965), both can be considered a slightly different forms of surface variation, that the approach developed is specifically designed to abstract away. For example, '`director ejecutivo`', the Spanish term for '`Chief Executive Officer`' can be considered, like the term '`CEO`', as just another synonym. The task of translating terms between languages (and maybe especially Indo-European languages such as English, French, Spanish, Italian, German) is becoming increasingly feasible, such that while several years ago this would have been a daunting task, the difficulty with advanced machine translations is now substantially reduced (Koehn et al., 2007; Wu et al., 2016). Similarly, the ontologies are already designed to allow multiple sentence constructions to be captured (e.g., '`John is the CEO`' vs. '`the CEO is John`'); only modest adjustments are thus necessary to expand the semantic rules to interpret differing orderings in different languages. The

flexibility of the ontologies to represent information in a different language – in this case, Spanish – is illustrated in Figure F.2.

a)

```
{
"ORIGINAL_SENTENCE":"Desde marzo de 2010 hasta mayo de 2015, Nicolás ha sido director ejecutivo de AMT de
                     México.",
"BACKGROUND":{"PERSON_NAME":{"ORIGINAL":"Nicolás","FIRST_NAME":"NICOLÁS"}
"POSITIONS":[
  { "START_DATE":{"ORIGINAL":"marzo de 2010","YEAR":2010,"MONTH":3},
    "END_DATE":{"ORIGINAL":"mayo de 2015","YEAR":2004,"MONTH":5},
    "POSITION":{
       "ORIGINAL":"ha sido director ejecutivo de AMT de México",
       "JOB_TITLES":[{"ORIGINAL":"director ejecutivo","LEVEL":"CEO"}],
       "COMPANY":{"ORIGINAL":"AMT de México","CLEANED":"AMT DE MÉXICO"}
  }]
}
```


b)

```
{
  "ORACION_ORIGINAL":"Pedro recibió una licenciatura en ingeniería de la Universidad Madero Puebla",
  "ANTECEDENTES":{"NOBRE_PERSONA":{"ORIGINAL":"Pedro","PRIMER_NOBRE":"PEDRO"}
  "CALIFACACION":[
    {
       "EDUCATION_INSTITUTION":{"ORIGINAL":"Universidad Madero Puebla",
                                "UNIVERSIDAD":"UNIVERSIDAD MADERO PUEBLA"},
       "QUALIFICATION":{"ORIGINAL":"licenciatura en ingeniería",
                        "LEVEL":"LICENCIATURA","SUJETOS":["INGENIERIA"]}
    }
  ]
}
```


Figure F.2. Illustrations of Ontologies Populated in Spanish, with either a) English or b) Spanish Ontology Labels

Although it may be necessary to increase the flexibility of various sub-ontologies to incorporate country-specific variations in governance practices not discussed in the backgrounds of US managers (e.g., the separation between Management and Supervisory boards in two-tier structuring of Germany firms: Fiss and Zajac, 2004), the underlying ontologies should largely transcend the US context. Being able to systematically capture meaning in a standardized manner, irrespective of whether the researcher speaks the language, offers researchers possibilities to develop theory that inherently require rich, multi-lingual information. Thus, while the ability to capture a representation of meaning across languages may help extend social science research beyond its English-language focus (Henrich, Heine, and Norenzayan, 2010), it also offers the possibility to develop and explore qualitatively different theories, including how information evolves within and across different global contexts.

BIBLIOGRAPHY

Abadi, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, et al.
2016 "Tensorflow: A system for large-scale machine learning." 12th USENIX Symposium on Operating Systems Design and Implementation Vol. 16: 265–283.

Abbott, A.
1990 "A primer on sequence methods." Organization Science, 1: 375–392.

Abrahamson, E., and D. C. Hambrick
1997 "Attentional homogeneity in industries: The effect of discretion." Journal of Organizational Behavior, 18: 513–532.

Abrahamsson, P., O. Salo, J. Ronkainen, and J. Warsta
2017 "Agile software development methods: Review and analysis." ArXiv, 1709.08439.

Abrams, R.
2014 "Walmart Vice President Forced Out for Lying About Degree." The New York Times.

Aggarwal, C. C., and C. Zhai
2012 Mining Text Data. New York, NY: Springer Science & Business Media.

Allen, M. J., and W. M. Yen
1979 Introduction to Measurement Theory. Monterrey, CA: Brooks/Cole.

Allred, B. B., C. C. Snow, and R. E. Miles
1996 "Characteristics of managerial careers of the 21st century." Academy of Management Executive, 10: 17–27.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman
1990 "Basic local alignment search tool." Journal of Molecular Biology, 215: 403–410.

Anderson, P., and M. L. Tushman
1990 "Technological discontinuities and dominant designs: A cyclical model of technological change." Administrative Science Quarterly, 35: 604–633.

Arthaud-Day, M. L., S. T. Certo, C. M. Dalton, and D. R. Dalton
2006 "A changing of the guard: Executive and director turnover following corporate financial restatements." Academy of Management Journal, 49: 1119–1136.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives
2007 "DBpedia: A Nucleus for a Web of Open Data." In Aberer, K., et al., (ed.), ISWC/ASWC 2007 Vol. 4825: 722–735. Heidelberg, Germany: Springer.

Baldridge, J.
2005 The OpenNLP Project. opennlp.apache.org.

Ballinger, G. A., and J. J. Marcel
2010 "The use of an interim CEO during succession episodes and firm performance." Strategic Management Journal, 31: 262–283.

Bao, Y., and A. Datta
2014 "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures." Management Science, 60: 1371–1391.

Barley, S. R., and G. Kunda
1992 "Design and devotion: Surges of rational and normative ideologies of control in managerial discourse." Administrative Science Quarterly, 37: 363–399.

Barnett, W. P., and M. T. Hansen
1996 "The red queen in organizational evolution." Strategic Management Journal, 17: 139–157.

Baum, J. A. C., and H. J. Korn
1996 "Competitive dynamics of interfirm rivalry." Academy of Management Journal, 39: 255–291.

BBC News Labs
2016 "Topic Modelling on BBC data." BBC News Labs. Retrieved from http://bbcnewslabs.co.uk/2016/06/27/topicmodel-update/

Bednar, M. K., S. Boivie, and N. R. Prince
2013 "Burr under the saddle: How media coverage influences strategic change." Organization Science, 24: 910–925.

Benner, M. J.
2010 "Securities analysts and incumbent response to radical technological change: Evidence from digital photography and internet telephony." Organization Science, 21: 42–62.

Berger, P. L., and T. Luckmann
1967 The Social Construction of Reality: A Treatise in the Sociology of Knowledge. London, UK: Penguin.

Bernardi, R. A., D. F. Bean, and K. M. Weippert
2002 "Signaling gender diversity through annual report pictures: A research note on image management." Accounting, Auditing & Accountability Journal, 15: 609–616.

Bettis, R. A., and M. A. Hitt
1995 "The new competitive landscape." Strategic Management Journal, 16: 7–19.

Bettman, J. R., and B. A. Weitz

1983 “Attributions in the board room: Causal reasoning in corporate annual reports.” Administrative Science Quarterly, 28: 165–183.

Biega, J., E. Kuzey, and F. M. Suchanek
2013 “Inside YAGO2s: A transparent information extraction architecture.” Proceedings of the 22nd International Conference on World Wide Web: 325–328. New York, NY: Association for Computing Machinery.

Bird, S., and E. Loper
2004 “NLTK: The natural language toolkit.” Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions: 31.

Blackstone, A.
2012 Principles of Sociological Inquiry: Qualitative and Quantitative Methods. Nyak, NY: Flat World Knowledge.

Blei, D. M., A. Y. Ng, and M. I. Jordan
2003 “Latent dirichlet allocation.” Journal of Machine Learning Research, 3: 993–1022.

Boje, D. M.
1991 “The storytelling organization: A study of story performance in an office- supply firm.” Administrative Science Quarterly, 36: 106–126.

Bolman, L. G., and T. E. Deal
2017 Reframing Organizations: Artistry, Choice, and Leadership. John Wiley & Sons.

Bourdieu, P.
1984 Distinction: A Social Critique of the Judgement of Taste. Cambridge, MA: Harvard University Press.

Bower, G., and T. Trabasso
1963 “Studies in Mathematical Psychology.” In R. C. Atkinson (ed.), Concept Identification: 32–94. Stanford, CA: Stanford University Press.

Boyd, J. L., and R. K. Bresser
2008 “Performance implications of delayed competitive responses: evidence from the US retail industry.” Strategic Management Journal, 29: 1077–1096.

Brown, L. D., A. C. Call, M. B. Clement, and N. Y. Sharp
2015 “Inside the ‘black box’ of sell-side financial analysts.” Journal of Accounting Research, 53: 1–47.

Campbell, D. T., and J. C. Stanley
1963 Experimental and Quasi-Experimental Designs for Research. Boston, MA: Houghton Mifflin Company.

Cannella Jr., A. A., J.-H. Park, and H.-U. Lee
2008 "Top management team functional background diversity and firm performance: Examining the roles of team member colocation and environmental uncertainty." Academy of Management Journal, 51: 768–784.

Carlson, A., J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell
2010 "Toward an architecture for never-ending language learning." Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence: 1306–1313. Atlanta, GA: AAAI Press.

Carnabuci, G., E. Operti, and B. Kovács
2016 "The categorical imperative and structural reproduction: Dynamics of technological entry in the semiconductor industry." Organization Science, 26: 1734–1751.

Carroll, G. R., and A. Swaminathan
2000 "Why the microbrewery movement? Organizational dynamics of resource partitioning in the US brewing industry." American Journal of Sociology, 106: 715–762.

Caruana, R., and A. Niculescu-Mizil
2006 "An empirical comparison of supervised learning algorithms." Proceedings of the 23rd International Conference on Machine Learning: 161–168.

Cattani, G., S. Ferriani, and P. D. Allison
2014 "Insiders, outsiders, and the struggle for consecration in cultural fields a core-periphery perspective." American Sociological Review, 79: 258–281.

Chatterjee, A., and D. C. Hambrick
2007 "It's all about me: Narcissistic chief executive officers and their effects on company strategy and performance." Administrative Science Quarterly, 52: 351–386.

Chen, G., D. C. Hambrick, and T. G. Pollock
2008 "Puttin' on the ritz: pre-IPO: Enlistment of prestigious affiliates as deadline-induced remediation." Academy of Management Journal, 51: 954–975.

Chen, M.-J., and D. C. Hambrick
1995 "Speed, stealth, and selective attack: How small firms differ from large firms in competitive behavior." Academy of Management Journal, 38: 453–482.

Chklovski, T., and P. Pantel
2004 "Verbocean: Mining the web for fine-grained semantic verb relations." Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.

Chowdhury, G. G.
2003 "Natural language processing." Annual Review of Information Science and Technology, 37: 51–89.

Chreim, S.
2005 "The continuity–change duality in narrative texts of organizational identity." Journal of Management Studies, 42: 567–593.

Cohen, A. M., and W. R. Hersh
2005 "A survey of current work in biomedical text mining." Briefings in Bioinformatics, 6: 57–71.

Cohen, B. D., and T. J. Dean
2005 "Information asymmetry and investor valuation of IPOs: Top management team legitimacy as a capital market signal." Strategic Management Journal, 26: 683–690.

Cohen, L., A. Frazzini, and C. J. Malloy
2012 "Hiring cheerleaders: Board appointments of 'independent' directors." Management Science, 58: 1039–1058.

Cohen, W., P. Ravikumar, and S. Fienberg
2003 "A comparison of string metrics for matching names and records." Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web, 73–78.

Courtis, J. K.
1986 "An investigation into annual report readability and corporate risk-return relationships." Accounting and Business Research, 16: 285–294.

Crilly, D., M. Hansen, and M. Zollo
2016 "The grammar of decoupling: A cognitive-linguistic perspective on firms' sustainability claims and stakeholders' interpretation." Academy of Management Journal, 59: 705–729.

Crossland, C., Jinyong, Zyung, N. J. Hiller, and D. C. Hambrick
2014 "CEO career variety: Effects on firm-level strategic and social novelty." Academy of Management Journal, 57: 652–674.

Csaszar, F. A., and N. Siggelkow
2010 "How much to copy? Determinants of effective imitation breadth." Organization Science, 21: 661–676.

Cucerzan, S.
2007 "Large-scale named entity disambiguation based on Wikipedia data." Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

Curran, J. R., T. Murphy, and B. Scholz
2007 "Minimising semantic drift with mutual exclusion bootstrapping." Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics Vol. 6: 172–180.

Cyert, R. M., and J. G. March
1963 A Behavioral Theory of the Firm. Englewood Cliffs, NJ: Prentice-Hall.

Daily, C. M., D. R. Dalton, and A. A. Cannella
2003 "Corporate governance: Decades of dialogue and data." Academy of Management Review, 28: 371–382.

D'Aveni, R. A., and I. C. MacMillan
1990 "Crisis and the content of managerial communications: A study of the focus of attention of top managers in surviving and failing firms." Administrative Science Quarterly, 35: 634–657.

Davis, G. F.
1991 "Agents without principles? The spread of the poison pill through the intercorporate network." Administrative Science Quarterly, 36: 583–613.

Deephouse, D. L.
1999 "To be different, or to be the same? It's a question (and theory) of strategic balance." Strategic Management Journal, 20: 147–166.

Drucker, H., D. Wu, and V. N. Vapnik
1999 "Support vector machines for spam categorization." IEEE Transactions on Neural Networks, 10: 1048–1054.

Duggan, G. B., and S. J. Payne
2011 "Skim reading by satisficing: Evidence from eye tracking." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: 1141–1150. New York, NY: Association for Computing Machinery.

Eagly, A. H., and L. L. Carli
2007 Through the Labyrinth: The Truth about How Women Become Leaders. Boston, MA: Harvard Business School Press.

Edgar, R. C.
2010 "Search and clustering orders of magnitude faster than BLAST." Bioinformatics, 26: 2460–2461.

Eisenhardt, K. M.
1989 "Building theories from case study research." Academy of Management Review, 14: 532–550.

Eisenhardt, K. M., and M. E. Graebner
2007 "Theory building from cases: Opportunities and challenges." Academy of Management Journal, 50: 25–32.

Elsbach, K. D.

1994 "Managing organizational legitimacy in the California cattle industry: The construction and effectiveness of verbal accounts." Administrative Science Quarterly, 39: 57–88.
2006 Organizational Perception Management. Mahwah, NJ: Psychology Press.

Ercan, G., and I. Cicekli
2007 "Using lexical chains for keyword extraction." Information Processing & Management, 43: 1705–1714.

Ertug, G., T. Yogev, Y. G. Lee, and P. Hedström
2016 "The art of representation: How audience-specific reputations affect success in the contemporary art field." Academy of Management Journal, 59: 113–134.

Fairclough, N.
1992 Discourse and Social Change. Cambridge, MA: Polity Press.

Fama, E. F., and M. C. Jensen
1983 "Separation of ownership and control." Journal of Law and Economics, 26: 301–325.

Finkel, J. R., T. Grenager, and C. Manning
2005 "Incorporating non-local information into information extraction systems by Gibbs sampling." Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics: 363–370.

Finkelstein, S., and R. A. D'Aveni
1994 "CEO duality as a double-edged sword: How boards of directors balance entrenchment avoidance and unity of command." Academy of Management Journal, 37: 1079–1108.

Finkelstein, S., and D. C. Hambrick
1990 "Top-management-team tenure and organizational outcomes: The moderating role of managerial discretion." Administrative Science Quarterly, 35: 484–503.

Fiol, C. M.
2002 "Capitalizing on paradox: The role of language in transforming organizational identities." Organization Science, 13: 653–666.

Firth, J. R.
1957 "A synopsis of linguistic theory, 1930-1955." Studies in Linguistic Analysis, Philological Society, Oxford; Reprinted in Palmer, F. (Ed.) 1968 Selected Papers of J. R. Firth, Longman, Harlow.

Fiss, P. C., and P. M. Hirsch
2005 "The discourse of globalization: Framing and sensemaking of an emerging concept." American Sociological Review, 70: 29–52.

Fiss, P. C., and E. J. Zajac

2004 “The diffusion of ideas over contested terrain: The (non) adoption of a shareholder value orientation among German firms.” Administrative Science Quarterly, 49: 501–534.
2006 “The symbolic management of strategic change: Sensegiving via framing and decoupling.” Academy of Management Journal, 49: 1173–1193.

Fleming, L., and O. Sorenson
2001 “Technology as a complex adaptive system: evidence from patent data.” Research Policy, 30: 1019–1039.

Franzosi, R.
1998 “Narrative analysis—or why (and how) sociologists should be interested in narrative.” Annual Review of Sociology, 24: 517–554.

Fredrickson, J. W., D. C. Hambrick, and S. Baumrin
1988 “A model of CEO dismissal.” Academy of Management Review, 13: 255–270.

Funk, R. J., and D. Hirschman
2014 “Derivatives and deregulation financial innovation and the demise of Glass–Steagall.” Administrative Science Quarterly, 59: 669–704.

Gamson, W. A., and A. Modigliani
1989 “Media discourse and public opinion on nuclear power: A constructionist approach.” American Journal of Sociology, 95: 1–37.

Geletkanycz, M. A., and D. C. Hambrick
1997 “The external ties of top executives: Implications for strategic choice and performance.” Administrative Science Quarterly, 42: 654–681.

Gensch, D. H.
1987 “A two-stage disaggregate attribute choice model.” Marketing Science, 6: 223–239.

Gioia, D. A., and K. Chittipeddi
1991 “Sensemaking and sensegiving in strategic change initiation.” Strategic Management Journal, 12: 433–448.

Glaser, B., and A. Strauss
1967 The Discovery of Grounded Theory: Strategies for Qualitative Research. New Brunswick: Aldine Transaction.

Goffman, E.
1959 The Presentation of Self in Everyday Life. New York: Penguin Harmondsworth.

Gomulya, D., and W. Boeker
2016 “Reassessing board member allegiance: CEO replacement following financial misconduct.” Strategic Management Journal, 37: 1898–1918.

Google
2003, September 24 "Serving advertisements based on content."
2012 "Introducing the Knowledge Graph: things, not strings." Official Google Blog. Retrieved from https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
2018 "Geocoding API." Google Developers. Retrieved from https://developers.google.com/maps/documentation/geocoding/start

Grimmer, J., and G. King
2011 "General purpose computer-assisted clustering and conceptualization." Proceedings of the National Academy of Sciences, 108: 2643–2650.

Gupta, V., and G. S. Lehal
2009 "A survey of text mining techniques and applications." Journal of Emerging Technologies in Web Intelligence, 1: 60–76.

Hambrick, D. C.
2005 "Upper echelons theory: Origins, twists and turns, and lessons learned." In K. G. Smith and M. A. Hitt (eds.), Great Minds in Management: The Process of Theory Development: 109–127. Oxford, UK: Oxofrd University Press.

Hambrick, D. C., T. S. Cho, and M. J. Chen
1996 "The influence of top management team heterogeneity on firms' competitive moves." Administrative Science Quarterly, 41: 659–684.

Hambrick, D. C., and P. A. Mason
1984 "Upper echelons: The organization as a reflection of its top managers." Academy of Management Review, 9: 193–206.

Han, J., J. Pei, and M. Kamber
2012 Data Mining: Concepts and Techniques. Boston, MA: Elsevier.

Hannigan, T.
2015 "Close encounters of the conceptual kind: Disambiguating social structure from text." Big Data & Society, 2: 1–6.

Harmon, D.
2018 "When the Fed speaks: Arguments, emotions, and the micro-foundations of institutions." Administrative Science Quarterly, Forthcoming.

Harmon, D. J., J. Green Sandy E., and G. T. Goodnight
2015 "A model of rhetorical legitimation: The structure of communication and cognition underlying institutional maintenance and change." Academy of Management Review, 40: 76–95.

Haunschild, P. R.

1993 “Interorganizational imitation: The impact of interlocks on corporate acquisition activity.” Administrative Science Quarterly, 38: 564–592.

Henrich, J., S. J. Heine, and A. Norenzayan
2010 “The weirdest people in the world?” Behavioral and Brain Sciences, 33: 61–83.

Hiatt, S. R., and Sangchan Park
2013 “Lords of the harvest: Third-party influence and regulatory approval of genetically modified organisms.” Academy of Management Journal, 56: 923–944.

Higgins, M. C., and R. Gulati
2003 “Getting off to a good start: The effects of upper echelon affiliations on underwriter prestige.” Organization Science, 14: 244–263.

Hipo
2015 University Domains List. Retrieved from https://github.com/Hipo/university-domains-list

Hirsch, P. M.
1986 “From Ambushes to Golden Parachutes: Corporate Takeovers as an Instance of Cultural Framing and Institutional Integration.” American Journal of Sociology, 91: 800–837.

Hoch, S. J., and G. F. Loewenstein
1991 “Time-inconsistent Preferences and Consumer Self-Control.” Journal of Consumer Research, 17: 492–507.

Hochreiter, S., and J. Schmidhuber
1997 “Long short-term memory.” Neural Computation, 9: 1735–1780.

Hochschild, A. R.
1983 The Managed Heart: Commercialization of Human Feeling. Berkeley, CA: University of California Press.

Hofmann, T.
1999 “Probabilistic latent semantic indexing.” Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: 50–57.

Hsu, G., and S. Grodal
2015 “Category taken-for-grantedness as a strategic opportunity the case of light cigarettes, 1964 to 1993.” American Sociological Review, 80: 28–62.

Hulth, A.
2003 “Improved automatic keyword extraction given more linguistic knowledge.” Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing: 216–223. Association for Computational Linguistics.

Jensen, M., and B. K. Kim
2014 "Great, Madame Butterfly again! How robust market identity shapes opera repertoires." Orangization Science, 25: 109–126.

Jensen, M., H. Kim, and B. K. Kim
2012 "Meeting expectations: A role-theoretic perspective on reputation." In M. L. Barnett and T. G. Pollock (eds.), The Oxford Handbook of Corporate Reputation. Oxford: Oxford University Press.

Jensen, M., and A. Roy
2008 "Staging exchange partner choices: When do status and reputation matter?" Academy of Management Journal, 51: 495–516.

Jurafsky, D., and J. H. Martin
2018 Speech and Language Processing (3rd edition: draft).

Kahl, S. J., and S. Grodal
2016 "Discursive strategies and radical technological change: Multilevel discourse analysis of the early computer (1947–1958)." Strategic Management Journal, 37: 149–166.

Kanhabua, N., R. Blanco, and M. Matthews
2011 "Ranking related news predictions." Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval: 755–764.

Kaplan, S., and K. Vakili
2015 "The double-edged sword of recombination in breakthrough innovation." Strategic Management Journal, 36: 1435–1457.

Kendall, S., and D. Tannen
1997 "Gender and language in the workplace." Gender and Discourse, 81–105.

Kennedy, M. T.
2005 "Behind the one-way mirror: Refraction in the construction of product market categories." Poetics, 33: 201–226.

Kim, B. K., and M. Jensen
2011 "How product order affects market identity: Repertoire ordering in the US opera market." Administrative Science Quarterly, 56: 238–256.

King, B. G.
2008 "A political mediation model of corporate response to social movement activism." Administrative Science Quarterly, 53: 395–421.

Kintsch, W., and T. A. van Dijk

1978 “Toward a model of text comprehension and production.” Psychological Review, 85: 363–394.

Kittur, A., E. H. Chi, and B. Suh
2008 “Crowdsourcing user studies with Mechanical Turk.” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: 453–456. New York, NY: Association for Computing Machinery.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, et al.
2007 “Moses: Open source toolkit for statistical machine translation.” Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions: 177–180. Stroudsburg, PA: Association for Computational Linguistics.

König, A., J. Mammen, J. Luger, A. Fehn, and A. Enders
2018 “Silver bullet or ricochet? CEOs’ use of metaphorical communication and infomediaries’ evaluations.” Academy of Management Journal, Forthcoming.

Kovács, B., and A. Sharkey
2013 “The paradox of publicity: How awards can negatively impact the evaluation of quality.” Administrative Science Quarterly, 59: 1–33.

Krause, R., R. Priem, and L. Love
2015 “Who’s in charge here? Co-CEOs, power gaps, and firm performance.” Strategic Management Journal, 36: 2099–2110.

Lacey, A., J. Lyons, A. Akbari, S. L. Turner, A. M. Walters, B. Fonferko-Shadrach, O. Pickrell, et al.
2017 “Codifying unstructured data: A Natural Language Processing approach to extract rich data from clinical letters.” International Journal for Population Data Science, 1.

Lakoff, G.
1975 “Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts.” Contemporary Research in Philosophical Logic and Linguistic Semantics, The University of Western Ontario Series in Philosophy of Science: 221–271. Springer, Dordrecht.

Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer
2016 “Neural architectures for named entity recognition.” ArXiv, 1603.01360.

Lawrence, T. B., and R. Suddaby
2006 “Institutions and institutional work.” In S. R. Clegg, T. Hardy, T. B. Lawrence, and W. R. Nord (eds.), The Sage Handbook of Organization Studies: 213–254. London, UK: SAGE.

Leung, M. D.
2014 “Dilettante or renaissance person? How the order of job experiences affects hiring in an external labor market.” American Sociological Review, 79: 136–158.

Levinthal, D. A.
1997 "Adaptation on rugged landscapes." Management Science, 43: 934–950.

Li, F.
2008 "Annual report readability, current earnings, and earnings persistence." Journal of Accounting and Economics, Economic Consequences of Alternative Accounting Standards and Regulation, 45: 221–247.

Lieberman, M. B., and D. B. Montgomery
1988 "First-mover advantages." Strategic Management Journal, 9: 41–58.

Livengood, R. S., and R. K. Reger
2010 "That's our turf! Identity domains and competitive dynamics." Academy of Management Review, 35: 48–66.

Loewenstein, J., W. Ocasio, and C. Jones
2012 "Vocabularies and vocabulary structure: A new approach linking categories, practices, and institutions." Academy of Management Annals, 6: 41–86.

Loughran, T., and B. McDonald
2011 "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." The Journal of Finance, 66: 35–65.

Magerman, T., B. V. Looy, and X. Song
2010 "Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications." Scientometrics, 82: 289–306.

Maguire, S., and C. Hardy
2009 "Discourse and deinstitutionalization: The decline of DDT." Academy of Management Journal, 52: 148–178.

Manrai, A. K., and R. L. Andrews
1998 "Two-stage discrete choice models for scanner panel data: An assessment of process and assumptions." European Journal of Operational Research, 111: 193–215.

Marcel, J. J.
2009 "Why top management team characteristics matter when employing a chief operating officer: A strategic contingency perspective." Strategic Management Journal, 30: 647–658.

March, J. G., and J. P. Olsen
1984 "The new institutionalism: Organizational factors in political life." The American Political Science Review, 78: 734–749.

Martens, M. L., J. E. Jennings, and P. D. Jennings

2007 “Do the stories they tell get them the money they need? The role of entrepreneurial narratives in resource acquisition.” Academy of Management Journal, 50: 1107–1132.

Martin, J., M. S. Feldman, M. J. Hatch, and S. B. Sitkin
1983 “The uniqueness paradox in organizational stories.” Administrative Science Quarterly, 28: 438–453.

Matsuo, Y., and M. Ishizuka
2004 “Keyword extraction from a single document using word co-occurrence statistical information.” International Journal on Artificial Intelligence Tools, 13: 157–169.

Matthews, P. H., and P. H. Matthews
1981 Syntax. Cambridge, UK: Cambridge University Press.

McGraw, K. M., M. Lodge, and P. Stroh
1990 “On-line processing in candidate evaluation: The effects of issue order, issue importance, and sophistication.” Political Behavior, 12: 41–58.

McGrimmon, L.
2014 The Resume Writing Guide: A Step-by-Step Workbook for Writing a Winning Resume 2 edition. Career Choice Guide.

Merton, R. K.
1957 Social Theory and Social Structure. Glencoe, IL: The Free Press.

Meyer, J. W., and B. Rowan
1977 “Institutionalized organizations: Formal structure as myth and ceremony.” American Journal of Sociology, 83: 340–363.

Michel, J. G., and D. C. Hambrick
1992 “Diversification posture and top management team characteristics.” Academy of Management Journal, 35: 9–37.

Munir, K. A., and N. Phillips
2005 “The birth of the ‘Kodak moment’: Institutional entrepreneurship and the adoption of new technologies.” Organization Studies, 26: 1665–1687.

Murphy, G. L.
2002 The Big Book of Concepts. Cambridge, MA: The MIT Press.

Nakashole, N., M. Theobald, and G. Weikum
2011 “Scalable knowledge harvesting with high precision and high recall.” Proceedings of the Fourth ACM International Conference on Web Search and Data Mining: 227–236. ACM.

Navis, C., and M. A. Glynn

2010 "How new market categories emerge: Temporal dynamics of legitimacy, identity, and entrepreneurship in satellite radio, 1990–2005." Administrative Science Quarterly, 55: 439–471.

Nelson, R. R., and S. G. Winter
1982 An Evolutionary Theory of Economic Change. Cambridge, MA: Belknap Press.

Nordgren, L. F., and A. Dijksterhuis
2009 "The devil is in the deliberation: Thinking too much reduces preference consistency." Journal of Consumer Research, 36: 39–46.

Ocasio, W., and J. Joseph
2005 "Cultural adaptation and institutional change: The evolution of vocabularies of corporate governance, 1972–2003." Poetics, 33: 163–178.

Ocasio, W., and H. Kim
1999 "The circulation of corporate control: Selection of functional backgrounds of new CEOs in large U.S. manufacturing firms, 1981–1992." Administrative Science Quarterly, 44: 532–562.

O'Reilly III, C. A., D. F. Caldwell, and W. P. Barnett
1989 "Work group demography, social integration, and turnover." Administrative Science Quarterly, 21–37.

Pang, B., L. Lee, and S. Vaithyanathan
2002 "Thumbs up?: Sentiment classification using machine learning techniques." Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing Vol. 10: 79–86. Association for Computational Linguistics.

Papadimitriou, C. H., P. Raghavan, H. Tamaki, and S. Vempala
2000 "Latent Semantic Indexing: A probabilistic analysis." Journal of Computer and System Sciences, 61: 217–235.

Perfetti, C. A., and L. Hart
2001 "The lexical basis of comprehension skill." On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity, Decade of behavior: 67–86. Washington, DC: American Psychological Association.

Pfarrer, M. D., K. A. Decelles, K. G. Smith, and M. S. Taylor
2008 "After the fall: Reintegrating the corrupt organization." Academy of Management Review, 33: 730–749.

Pfeffer, J.
1981 Power in Organizations. Marshfield, MA: Pitman.

Philipsen, G.
1975 "Speaking 'like a man' in Teamsterville: Culture patterns of role enactment in an urban neighborhood." Quarterly Journal of Speech, 61: 13.

Phillips, D. J., and Y.-K. Kim
2009 "Why pseudonyms? Deception as identity preservation among jazz record companies, 1920-1929." Organization Science, 20: 481–499.

Phillips, D. J., and E. W. Zuckerman
2001 "Middle-status conformity: Theoretical restatement and empirical demonstration in two markets." American Journal of Sociology, 107: 379–429.

Potter, J.
1996 Representing Reality: Discourse, Rhetoric and Social Construction. Thousand Oaks, CA: SAGE.

Pritchard, J. K., M. Stephens, and P. Donnelly
2000 "Inference of population structure using multilocus genotype data." Genetics, 155: 945–959.

Radner, R.
1992 "Hierarchy: The economics of managing." Journal of Economic Literature, 30: 1382–1415.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang
2016 "SQuAD: 100,000+ questions for machine comprehension of text." ArXiv, 1606.05250.

Reger, R. K., L. T. Gustafson, S. M. Demarie, and J. V. Mullane
1994 "Reframing the organization: Why implementing total quality is easier said than done." Academy of Management Review, 19: 565–584.

Rennekamp, K.
2012 "Processing fluency and investors' reactions to disclosure readability." Journal of Accounting Research, 50: 1319–1354.

Rhee, E. Y., and P. C. Fiss
2014 "Framing controversial actions: Regulatory focus, source credibility, and stock market reaction to poison pill adoption." Academy of Management Journal, 57: 1734–1758.

Riloff, E., and R. Jones
1999 "Learning dictionaries for information extraction by multi-level bootstrapping." Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference: 474–479. Menlo Park, CA: American Association for Artificial Intelligence.

Rivkin, J. W.
2000 "Imitation of complex strategies." Management Science, 46: 824–844.

Rosch, E. H.
1973 "Natural categories." Cognitive Psychology, 4: 328–350.

Rose, S., D. Engel, N. Cramer, and W. Cowley
2010 "Automatic keyword extraction from individual documents." Text Mining: Applications and Theory, 1–20.

Ryan, G. W., and H. R. Bernard
2003 "Techniques to identify themes." Field Methods, 15: 85–109.

Salancik, G. R., and J. R. Meindl
1984 "Corporate attributions as strategic illusions of management control." Administrative Science Quarterly, 29: 238–254.

Scott, M. B., and S. M. Lyman
1968 "Accounts." American Sociological Review, 33: 46–62.

Sebastiani, F.
2002 "Machine Learning in Automated Text Categorization." ACM Computing Surveys, 34: 1–47.

Seo, M.-G., and W. E. D. Creed
2002 "Institutional contradictions, praxis, and institutional change: A dialectical perspective." Academy of Management Review, 27: 222–247.

Sillince, J., P. Jarzabkowski, and D. Shaw
2012 "Shaping strategic action through the rhetorical construction and exploitation of ambiguity." Organization Science, 23: 630–650.

Simons, T., L. H. Pelled, and K. A. Smith
1999 "Making use of difference: diversity, debate, and decision comprehensiveness in top management teams." Academy of Management Journal, 42: 662–673.

Simpson, M. S., and D. Demner-Fushman
2012 "Biomedical text mining: A survey of recent progress." Mining Text Data: 465–517. Springer, Boston, MA.

Sonenshein, S.
2007 "The role of construction, intuition, and justification in responding to ethical issues at work: the sensemaking-intuition model." Academy of Management Review, 32: 1022–1040.

Specht, D. F.
1991 "A general regression neural network." IEEE Transactions on Neural Networks, 2: 568–576.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov
2014 "Dropout: A simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research, 15: 1929–1958.

Staw, B. M., P. I. McKechnie, and S. M. Puffer
1983 "The justification of organizational performance." Administrative Science Quarterly, 28: 582–600.

Stockwell, R. P., J. D. Bowen, and J. W. Martin
1965 The Grammatical Structures of English and Spanish. University of Chicago Press.

Strauss, A., and J. Corbin
1990 Basics of Qualitative Research Vol. 15. Newbury Park, CA: Sage.

Stuart, T. E., and J. M. Podolny
1996 "Local search and the evolution of technological capabilities." Strategic Management Journal, 17: 21–38.

study.com
2018 "Types of Different Degree Levels." Study.Com. Retrieved from http://study.com/different_degrees.html

Suchanek, F. M., G. Kasneci, and G. Weikum
2007 "Yago: A Core of Semantic Knowledge." Proceedings of the 16th International Conference on World Wide Web: 697–706. New York, NY: Association for Computing Machinery.

Suchanek, F. M., M. Sozio, and G. Weikum
2009 "SOFIE: A self-organizing framework for information extraction." Proceedings of the 18th International Conference on World Wide Web: 631–640. New York, NY: Association for Computing Machinery.

Sutton, R. I., and A. L. Callahan
1987 "The stigma of bankruptcy: Spoiled organizational image and its management." Academy of Management Journal, 30: 405–436.

Swinney, D. A.
1979 "Lexical access during sentence comprehension: (Re)consideration of context effects." Journal of Verbal Learning and Verbal Behavior, 18: 645–659.

Tausczik, Y. R., and J. W. Pennebaker
2010 "The psychological meaning of words: LIWC and computerized text analysis methods." Journal of Language and Social Psychology, 29: 24–54.

The Constant Analyst

2013 “CPA vs CMA vs CFA.” The Constant Analyst. Retrieved from http://theconstantanalyst.com/blog/2013/01/cpa-vs-cma-vs-cfa

Tripsas, M., and G. Gavetti
2000 “Capabilities, cognition, and inertia: Evidence from digital imaging.” Strategic Management Journal, 21: 1147–1161.

Tsai, W., K. Su, and M.-J. Chen
2011 “Seeing through the eyes of a rival: Competitor acumen based on rival-centric perceptions.” Academy of Management Journal, 54: 761–778.

Turney, P. D.
2008 “The latent relation mapping engine: Algorithm and experiments.” Journal of Artificial Intelligence Research, 33: 615–655.

Turney, P. D., and P. Pantel
2010 “From frequency to meaning: Vector space models of semantics.” Journal of Artificial Intelligence Research, 37: 141–188.

Tversky, A., and D. Kahneman
1981 “The framing of decisions and the psychology of choice.” Science, 211: 453–458.

Van Valin, R. D., and R. J. LaPolla
1997 Syntax: Structure, Meaning, and Function. Cambridge University Press.

Wade, J. B., J. F. Porac, and T. G. Pollock
1997 “Worth, words, and the justification of executive pay.” Journal of Organizational Behavior, 18: 641–664.

Weber, R.
1985 Basic Content Analysis. London, UK: Sage.

Westphal, J. D., and M. E. Graebner
2010 “A matter of appearances: How corporate leaders manage the impressions of financial analysts about the conduct of their boards.” Academy of Management Journal, 53: 15–44.

Westphal, J. D., R. Gulati, and S. M. Shortell
1997 “Customization or conformity? An institutional and network perspective on the content and consequences of TQM adoption.” Administrative Science Quarterly, 42: 366–394.

Westphal, J. D., and E. J. Zajac
2001 “Decoupling policy from practice: The case of stock repurchase programs.” Administrative Science Quarterly, 46: 202–228.

Wiesenfeld, B. M., K. A. Wurthmann, and D. C. Hambrick

2008 “The stigmatization and devaluation of elites associated with corporate failures: A process model.” Academy of Management Review, 33: 231–251.

Wilson, A. J., and J. Joseph
2015 “Organizational attention and technological search in the multibusiness firm: Motorola from 1974 to 1997.” Cognition and Strategy, Advances in Strategic Management Vol. 32: 407–435. Bingley, UK: Emerald Group.

Wooldridge, B., and S. W. Floyd
1990 “The strategy process, middle management, and organizational performance.” Strategic Management Journal, 11: 231–241.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, et al.
2016 “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.” ArXiv, 1609.08144.

Yoon, B., and Y. Park
2004 “A text-mining-based patent network: Analytical tool for high-technology trend.” The Journal of High Technology Management Research, 15: 37–50.

Zajac, E. J., and J. D. Westphal
2004 “The social construction of market value: Institutionalization and learning perspectives on stock market reactions.” American Sociological Review, 69: 433–457.

Zhang, Y., and M. F. Wiersema
2009 “Stock market reaction to CEO certification: The signaling role of CEO background.” Strategic Management Journal, 30: 693–710.

Zott, C., and Q. N. Huy
2007 “How entrepreneurs use symbolic management to acquire resources.” Administrative Science Quarterly, 52: 70–105.

Zwick, R., E. Caristein, and D. V. Budescu
1987 “Measures of Similarity Among Fuzzy Concepts: A Comparative Analysis.” International Journal of Approximate Reasoning, 1: 221–242.