Sage Research Methods

An Introduction to Text Mining: Research Design, Data Collection, and Analysis

For the most optimal reading experience we recommend using our website.

https://methods.sagepub.com/book/mono/an-introduction-to-text-mining/toc

Author: Gabe Ignatow, Rada Mihalcea

Pub. Date: 2021

Product: Sage Research Methods

DOI: https://doi.org/10.4135/9781506336985

Methods: Text mining, Social network research, Narrative research

Keywords: language, mining, software, social science, emotion, sociology, decision making

Disciplines: Business and Management, Criminology and Criminal Justice, Communication and Media Studies, Marketing, Political Science and International Relations, Sociology

Access Date: February 17, 2025

Publisher: SAGE Publications, Inc
City: Thousand Oaks
Online ISBN: 9781071849361

Front Matter

Chapters

Back Matter

Copyright

None

FOR INFORMATION:

SAGE Publications, Inc.

2455 Teller Road

Thousand Oaks, California 91320

E-mail: order@sagepub.com

SAGE Publications Ltd.

1 Oliver’s Yard

55 City Road

London, EC1Y 1SP

United Kingdom

SAGE Publications India Pvt. Ltd.

B 1/I 1 Mohan Cooperative Industrial Area

Mathura Road, New Delhi 110 044

India

SAGE Publications Asia-Pacific Pte. Ltd.

3 Church Street

#10-04 Samsung Hub

Singapore 049483

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Names: Ignatow, Gabe, author. | Mihalcea, Rada, 1974- author.

Title: An introduction to text mining : research design, data collection, and analysis / Gabe Ignatow, University of North Texas, Rada Mihalcea, University of Michigan.

Description: Thousand Oaks : SAGE Publications, [2018] | Includes bibliographical references and index.

Identifiers: LCCN 2017038203 | ISBN 9781506337005 (pbk. : alk. paper)

Subjects: LCSH: Data mining. | Social sciences—Research.

Classification: LCC QA76.9.D343 I425 2017 | DDC 006.3/12—dc23 LC record available at https://lccn.loc.gov/2017038203

This book is printed on acid-free paper.

Acquisitions Editor: Helen Salmon

Editorial Assistant: Megan O’Heffernan

eLearning Editor: Chelsea Neve

Production Editor: Kelly DeRosa

Copy Editor: Megan Markanich

Typesetter: C&M Digitals (P) Ltd.

Proofreader: Wendy Jo Dymond

Indexer: Joan Shapiro

Cover Designer: Michael Dubowe

Marketing Manager: Shari Countryman

Acknowledgments

An Introduction to Text Mining has been a long time in the making, and there are too many people to count who deserve our thanks for helping to bring this book to publication. First and foremost, we must thank our undergraduate and graduate students who have shown so much enthusiasm for learning about online communities. It was their energy and questions that convinced us of the need for this book. Helen Salmon, Katie Ancheta, and the entire editorial and production staff at SAGE deserve our special thanks. In truth, it was Helen who got this project off the ground, and she and the entire SAGE staff, including SAGE's team of expert reviewers, provided support and guidance throughout the writing and production process. SAGE's reviewers played an especially critical role by providing invaluable feedback based on their research and teaching experiences in their home disciplines. A textbook as interdisciplinary as this one requires absolutely top-flight reviewers, and we were fortunate to have many of them. A special thank-you goes to Roger Clark, Kate de Medeiros, Carol Ann MacGregor, Kenneth C. C. Yang, A. Victor Ferreros, and Jennifer Bachner.

Last but by no means least we thank our spouses and children Neva, Alex, and Sara, and Mihai, Zara, and Caius, for their patience with us and their encouragement over the many years of research, writing, and editing that went into this textbook.

GI and RM

Preface

Students are accustomed to participating in all sorts of online communities. While interacting on platforms such as Facebook, Twitter, Snapchat, and Instagram as well as on blogs, forums, and many other apps and sites, some students taking courses in the social sciences and computer science want to take things a step further and perform their own research on the social interactions that occur in these communities. We have written this book for those students, including especially undergraduate and graduate students in anthropology, communications, computer science, education, linguistics, marketing, political science, psychology, and sociology courses who want to do research using online tools and data sets. Whether they are writing a term paper or honors thesis, or working on an independent research project or a project with a faculty adviser, students who want to use text mining tools for social research need a place to start.

Online communities offer no end of interesting linguistic and social material to study, from emojis and abbreviations to forms of address, themes, metaphors, and all sorts of interpersonal conversational dynamics. The volume of data available for research, and the many research tools available to students, are simply overwhelming. *An Introduction to Text Mining* is here to help. The book is organized to guide students through major ethical, philosophical, and logical issues that should be considered in the earliest stages of a research project (see Part I) and then to survey the landscape of text mining and text analysis tools and methodologies that have been developed across the social sciences and computational linguistics. Appendices A through G on data and software resources are a key to the book, and readers should consider reviewing these early and returning to them often as they work their way through the early chapters and begin to design their own research projects (see Chapter 5).

If you think of your text mining research project as a house, then the chapters in Part I are instructions for building the foundation. Just as a house with a flaw in its foundation will not last long, a research project with a shaky logical foundation or questionable ethics may look good at the start, but it is inevitable that at some point its flaws will be exposed. Chapter 5 on research design provides architectural instruction for building the framework of your house. Designing a research project that can address, and perhaps conclusively answer, a research question or questions is a challenging task, and it is useful to know the kinds of research designs that have a track record of success in research using text mining tools and methodologies. Parts III through V survey text mining and analysis methodologies, the equivalent of proven house-building methods. Appendix A provides a partial survey of online sources of textual data, which is the raw material of your research project. Appendices B through G provide, as it were, a survey of the practical tools that are available for house construction, from hand tools to heavy-duty machinery. While setting the foundation, designing the house, and

choosing a construction method, it is a good idea to be aware of the types of practical tools that are available and within budget so that your project can reach a successful conclusion. Appendices H and I, as well as the Glossary, provide handy summaries of web resources, statistical tools, and key terms.

Additional resources for instructors using *An Introduction to Text Mining* are also provided. Editable, chapter-specific Microsoft¯ PowerPoint¯ slides, as well as assignments and activities created by the authors, are available for download at: **http:/study.sagepub.com/introtextmining**.

Note to the Reader

An Introduction to Text Mining grew out of our earlier SAGE methods guidebook *Text Mining*, which is a shorter volume intended to serve as a practical guidebook for graduate students and professional researchers. The two books share both a core mission and structure. Their mission is to enable readers to make better informed decisions about research projects that use text mining and text analysis methodologies. And they both survey text mining tools developed in multiple disciplines within the social sciences, humanities, and computer science.

Where *Text Mining* was intended for advanced students and researchers, the current volume is a dedicated undergraduate or first-year graduate textbook intended for use in social science and data science courses. This book is thus longer than *Text Mining*, as it includes new material related to ethical and epistemological considerations in text-based research. There is a new chapter on how to write text-based social science research papers. And there are appendices that list and review data sources and software for preparing, cleaning, organizing, analyzing, and visualizing patterns in texts. Although these appendices were intended for students in undergraduate courses we suspect that they will prove valuable for experienced researchers as well.

GI and RM

About the Authors

Gabe Ignatow is an associate professor of sociology at the University of North Texas (UNT), where he has taught since 2007. His research interests are in the areas of sociological theory, text mining and analysis methods, new media, and information policy. Gabe's current research involves working with computer scientists and statisticians to adapt text mining and topic modeling techniques for social science applications. Gabe has been working with mixed methods of text analysis since the 1990s and has published this work in the following journals: *Social Forces*, *Sociological Forum*, *Poetics*, the *Journal for the Theory of Social Behaviour*, and the *Journal of Computer-Mediated Communication*. He is the author of over 30 peer-reviewed articles and book chapters and serves on the editorial boards of the journals *Sociological Forum*, the *Journal for the Theory of Social Behaviour*, and *Studies in Media and Communication*. He has served as the UNT Department of Sociology's graduate program codirector and undergraduate program director and has been selected as a faculty fellow at the Center for Cultural Sociology at Yale University. He is also a cofounder and the CEO of GradTrek, a graduate degree search engine company.

Rada Mihalcea is a professor of computer science and engineering at the University of Michigan. Her research interests are in computational linguistics, with a focus on lexical semantics, multilingual natural language processing, and computational social sciences. She serves or has served on the editorial boards of the following journals: *Computational Linguistics*, *Language Resources and Evaluation*, *Natural Language Engineering*, *Research on Language and Computation*, *IEEE Transactions on Affective Computing*, and *Transactions of the Association for Computational Linguistics*. She was a general chair for the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL, 2015) and a program cochair for the Conference of the Association for Computational Linguistics (2011) and the Conference on Empirical Methods in Natural Language Processing (2009). She is the recipient of a National Science Foundation CAREER award (2008) and a Presidential Early Career Award for Scientists and Engineers (2009). In 2013, she was made an honorary citizen of her hometown of Cluj-Napoca, Romania.

Text Mining and Text Analysis

Learning Objectives

The goals of Chapter 1 are to help you to do the following:

1. Familiarize yourself with a variety of research projects accomplished using **text mining** tools.
2. Address different research questions using text mining tools.
3. Differentiate between text mining and **text analysis** methodologies.
4. Compare major theoretical and methodological approaches to both text mining and text analysis.

Introduction

Text mining is an exciting field that encompasses new research methods and software tools that are being used across academia as well as by companies and government agencies. Researchers today are using text mining tools in ambitious projects to attempt to predict everything from the direction of stock markets (Bollen, Mao, & Zeng, 2011) to the occurrence of political protests (Kallus, 2014). Text mining is also commonly used in marketing research and many other business applications as well as in government and defense work.

Over the past few years, text mining has started to catch on in the social sciences, in academic disciplines as diverse as anthropology (Acerbi, Lampos, Garnett, & Bentley, 2013; Marwick, 2013), communications (Lazard, Scheinfeld, Bernhardt, Wilcox, & Suran, 2015), economics (Levenberg, Pulman, Moilanen, Simpson, & Roberts, 2014), education (Evison, 2013), political science (Eshbaugh-Soha, 2010; Grimmer & Stewart, 2013), psychology (Colley & Neal, 2012; Schmitt, 2005), and sociology (Bail, 2012; Heritage & Raymond, 2005; Mische, 2014). Before social scientists began to adapt text mining tools to use in their research, they spent decades studying transcribed interviews, newspaper articles, speeches, and other forms of textual data, and they developed sophisticated text analysis methods that we review in the chapters in Part IV. So while text mining is a relatively new interdisciplinary field based in computer science, text analysis methods have a long history in the social sciences (see Roberts, 1997).

Text mining processes typically include information retrieval (methods for acquiring texts) and applications of

advanced statistical methods and **natural language processing (NLP)** such as part-of-speech tagging and syntactic parsing. Text mining also often involves named entity recognition (NER), which is the use of statistical techniques to identify named text features such as people, organizations, and place names; **disambiguation**, which is the use of contextual clues to decide where words refer to one or another of their multiple meanings; and **sentiment analysis**, which involves discerning subjective material and extracting attitudinal information such as sentiment, opinion, mood, and emotion. These techniques are covered in Parts III and V of this book. Text mining also involves more basic techniques for acquiring and processing data. These techniques include tools for **web scraping** and **web crawling**, for making use of dictionaries and other lexical resources, and for processing texts and relating words to texts. These techniques are covered in Parts II and III.

Research in the Spotlight

Predicting the Stock Market With Twitter

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

The computer scientists Bollen, Mao, and Zeng asked whether societies can experience mood states that affect their collective decision making, and by extension whether the public mood is correlated or even predictive of economic indicators. Applying sentiment analysis (see Chapter 14) to large-scale Twitter feeds, Bollen and colleagues investigated whether measurements of collective mood states are correlated to the value of the Dow Jones Industrial Average over time. They analyzed the text content of daily Twitter feeds using OpinionFinder, which measures positive versus negative mood and Google Profile of Mood States to measure mood in terms of six dimensions (calm, alert, sure, vital, kind, and happy). They also investigated the hypothesis that public mood states are predictive of changes in Dow Jones Industrial Average closing values, finding that the accuracy of stock market predictions can be significantly improved by the inclusion of some specific public mood dimensions but not others.

Specialized software used:

OpinionFinder

http://mpqa.cs.pitt.edu/opinionfinder

Text analysis involves systematic analysis of word use patterns in texts and typically combines formal statistical methods and less formal, more humanistic interpretive techniques. Text analysis arguably originated as early as the 1200s with the Dominican friar Hugh of Saint-Cher and his team of several hundred fellow friars who created the first biblical **concordance**, or cross-listing of terms and concepts in the Bible. There is also evidence of European inquisitorial church studies of newspapers in the late 1600s, and the first well-documented quantitative text analysis was performed in Sweden in the 1700s when the Swedish state church analyzed the symbology and ideological content of popular hymns that appeared to challenge church orthodoxy (Krippendorff, 2013, pp. 10–11). The field of text analysis expanded rapidly in the 20th century as researchers in the social sciences and humanities developed a broad spectrum of techniques for analyzing texts, including methods that relied heavily on human interpretation of texts as well as formal statistical methods. Systematic quantitative analysis of newspapers was performed in the late 1800s and early 1900s by researchers including Speed (1893), who showed that in the late 1800s New York newspapers had decreased their coverage of literary, scientific, and religious matters in favor of sports, gossip, and scandals. Similar text analysis studies were performed by Wilcox (1900), Fenton (1911), and White (1924), all of whom quantified newspaper space devoted to different categories of news. In the 1920s through 1940s, Lasswell and his colleagues conducted breakthrough **content analysis** studies of political messages and propaganda (e.g., Lasswell, 1927). Lasswell's work inspired large-scale content analysis projects including the **General Inquirer project** at Harvard, which is a lexicon attaching syntactic, semantic, and pragmatic information to part-of-speech tagged words (Stone, Dunphry, Smith, & Ogilvie, 1966).

While text mining's roots are in computer science and the roots of text analysis are in the social sciences and humanities, today, as we will see throughout this textbook, the two fields are converging. Social scientists and humanities scholars are adapting text mining tools for their research projects, while text mining specialists are investigating the kinds of social phenomena (e.g., political protests and other forms of collective behavior) that have traditionally been studied within the social sciences.

Six Approaches to Text Analysis

The field of text mining is divided mainly in terms of different methodologies, while the field of text analysis can be divided into several different approaches that are each based on a different way of theorizing language use. Before discussing some of the special challenges associated with using online data for social science research, next we review six of the most prominent approaches to text analysis. As we will see, many researchers who work with these approaches are finding ways to make use of the new text mining methodologies and tools that are covered in Parts II, III, and V. These approaches include **conversation analysis**, xe "analysis of discourse positions"analysis of **discourse positions**, **critical discourse analysis (CDA)**, content analysis, **Foucauldian analysis**, and analysis of texts as social information. These approaches use different logical strategies and are based on different theoretical foundations and philosophical assumptions (discussed in Chapter 4). They also operate at different levels of analysis (micro, meso, and macro) and employ different selection and sampling strategies (see Chapter 5).

Conversation Analysis

Conversation analysts study everyday conversations in terms of how people negotiate the meaning of the conversation in which they are participating and the larger discourse of which the conversation is a part. Conversation analysts focus not only on what is said in daily conversations but also on how people use language pragmatically to define the situations in which they find themselves. These processes go mostly unnoticed until there is disagreement as to the meaning of a particular situation. An example of conversation analysis is the educational researcher Evison's (2013) study of "academic talk," which used corpus linguistic techniques (see Appendix F) on both a corpus of 250,000 words of spoken academic discourse and a benchmark corpus of casual conversation to explore conversational turn openings. The corpus of academic discourse included 13,337 turns taken by tutors and students in a range of social interactions. In seeking to better understand the unique language of academia and of specific academic disciplines, Evison identified six items that have a particularly strong affinity with the turn-opening position (*mhm, mm, yes, laughter, oh, no*) as key characteristics of academic talk.

Further examples of conversation analysis research include studies of conversation in educational settings by O'Keefe and Walsh (2012); in health care settings by Heath and Luff (2000), Heritage and Raymond (2005), and Silverman (2016); and in online environments among Wikipedia editors by Danescu-Niculescu-Mizil, Lee,

Pang, and Kleinberg (2012). O'Keefe and Walsh's 2012 study combined corpus linguistics and conversation analysis methodologies to analyze higher education small-group teaching sessions. Their data are from a 1-million-word corpus, the Limerick–Belfast Corpus of Academic Spoken English (LIBEL CASE). Danescu-Niculescu-Mizil and colleagues (2012) analyzed signals manifested in language in order to learn about roles, status, and other aspects of groups' interactional dynamics. In their study of Wikipedians and of arguments before the U.S. Supreme Court, they showed that in group discussions, power differentials between participants are subtly revealed by the degree to which one individual immediately echoes the linguistic style of the person to whom they are responding. They proposed an analysis framework based on linguistic coordination that can be used to shed light on power relationships and that works consistently across multiple types of power, including more static forms of power based on status differences and more situational forms in which one individual experiences a type of dependence on another.

Hakimnia and her colleagues' (2015) conversation analysis of transcripts of calls to a telenursing site in Sweden used a comparative research design (see Chapter 5). The study's goal was to analyze callers' reasons for calling and the outcome of the calls in terms of whether men and women received different kinds of referrals. The researchers chose to randomly sample 800 calls from a corpus of over 5,000 total calls that had been recorded at a telenursing site in Sweden over a period of 11 months. Callers were informed about the study in a prerecorded message and consented to participate, while the nurses were informed verbally about the study. The first step in the analysis of the final sample of 800 calls was to create a matrix (see Chapter 5 and Appendices C and D), including information on each caller's gender, age, fluency or nonfluency in Swedish as well as the outcome of the call (whether callers were referred to a general practitioner). The researchers found that men, and especially fathers, received more referrals to general practitioners than did women. The most common caller was a woman fluent in Swedish (64%), and the least likely caller was a man nonfluent in Swedish (3%). All in all, 70% of the callers were women. When the calls concerned children, 78% of the callers were female. Based on these results, the researchers concluded that it is important that telenursing not become a "feminine" activity, only suitable for young callers fluent in Swedish. Given the telenurses' gatekeeping role, there is a risk that differences on this first level of health care could be reproduced throughout the whole health care system.

Analysis of Discourse Positions

Analyzing discourse positions is an approach to text analysis that allows researchers to reconstruct communicative interactions through which texts are produced and in this way gain a better understanding of their meaning from their author's viewpoint. Discourse positions are understood as typical discursive roles that people adopt in their everyday communication practices, and the analysis of discourse positions is a way of linking texts to the social spaces in which they have emerged. An example of contemporary discourse position research is Bamberg's (2004) study of the "small stories" told by adolescents and postadolescents about their identities. Bamberg's 2004 study is informed by theories of human development and of narrative (see Chapter 10). His texts are excerpts of transcriptions from a group discussion among five 15-year-old boys telling a story about a female student they all know. The group discussion was conducted in the presence of an adult moderator, but the data were collected as part of a larger project in which Bamberg and his colleagues collected journal entries and transcribed oral accounts from 10-, 12-, and 15-year-old boys in one-on-one interviews and group discussions. Although the interviews and groups discussions were open-ended, they all focused on the same list of topics, including friends and friendships, girls, the boys' feelings and sense of self, and their ideas about adulthood and future orientation. Bamberg and his team analyzed the transcripts line by line, coding instances of the boys positioning themselves relative to each other and to characters in their stories.

Edley and Wetherell's (1997, 2001; Wetherell & Edley, 1999) studies of masculine identity formation are similar to Bamberg's study in that they also focus on stories people tell themselves and others in ordinary everyday conversations. Edley and Wetherell studied a corpus of men's talk on feminism and feminists to identify patterns and regularities in their accounts of feminism and in the organization of their rhetoric. Their samples of men included a sample of white, middle-class 17- to 18-year-old school students and a sample of 60 interviews with a more diverse sample of older men aged 20 to 64. The researchers identified two "interpretative repertoires of feminism and feminists," which set up a "Jekyll and Hyde" binary and "positioned feminism along with feminists very differently as reasonable versus extreme" (Edley & Wetherell, 2001, p. 439).

In the end, analysis of discourse positions is for the most part a qualitative approach to text analysis that relies almost entirely on human interpretation of texts (see Hewson, 2014). Appendix D includes a list of contemporary qualitative data analysis software (QDAS) packages that can be used to organize and code the kinds of text corpora analyzed by Bamberg, Edley, Wetherell, and other researchers working in this tradition.

Critical Discourse Analysis

CDA involves seeking the presence of features from other discourses in the text or discourse to be analyzed. CDA is based on Fairclough's (1995) concept of "intertextuality," which is the idea that people appropriate from discourses circulating in their social space whenever they speak or write. In CDA, ordinary everyday speaking and writing are understood to involve selecting and combining elements from dominant discourses.

While the term *discourse* generally refers to all practices of writing and talking, in CDA discourses are understood as ways of writing and talking that "rule out" and "rule in" ways of constructing knowledge about topics. In other words, discourses "do not just describe things; they do things" (Potter & Wetherell, 1987, p. 6) through the way they make sense of the world for its inhabitants (Fairclough, 1992; van Dijk, 1993).

Discourses cannot be studied directly but can be explored by examining the texts that constitute them (Fairclough, 1992; Parker, 1992). In this way, texts can be analyzed as fragments of discourses that reflect and project ideological domination by powerful groups in society. But texts can also be considered a potential mechanism of liberation when they are produced by the critical analyst who reveals mechanisms of ideological domination in them in an attempt to overcome or eliminate them.

Although CDA has generally employed strictly interpretive methods, use of quantitative and statistical techniques is not a novel practice (Krishnamurthy, 1996; Stubbs, 1994), and the use of software to create, manage, and analyze large collections of texts appears to be increasingly popular (Baker et al., 2008; Koller & Mautner, 2004; O'Halloran & Coffin, 2004).

A 2014 study by Bednarek and Caple exemplifies the use of statistical techniques in CDA. Bednarek and Caple introduced the concept of "news values" to CDA of news media and illustrated their approach with two case studies using the same collection of British news discourse. Their texts included 100 news stories (about 70,000 words total) from 2003 covering 10 topics from 10 different national newspapers, including five quality papers and five tabloids. The analysis proceeded through analysis of word frequency of the top 100 most frequently used words and two-word clusters (bigrams), focusing on words that represent news values such as *eliteness, superlativeness, proximity, negativity, timeliness, personalization*, and *novelty*. The authors concluded that their case studies demonstrated that corpus linguistic techniques (see Appendix F) can identify discursive devices that are repeatedly used in news discourse to construct and perpetuate an ideology of newsworthiness.

In another CDA study, Baker and his colleagues (2008) analyzed a 140-million-word corpus of British news articles about refugees, asylum seekers, immigrants, and migrants. They used collocation and concordance analysis (see Appendix F) to identify common categories of representation of refugees, asylum seekers, immigrants, and migrants. They also discussed how collocation and concordance analysis can be used to direct researchers to representative texts in order to carry out qualitative analysis.

Research in the Spotlight

Combining Critical Discourse Analysis and Corpus Linguistics

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., Mcenery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273–306.

In this critical discourse analysis (CDA) study, the linguist Baker and his colleagues analyzed a 140-million-word corpus of British news articles about refugees, asylum seekers, immigrants, and migrants. The authors used collocation and concordance analysis (see Appendix F) to identify common categories of representations of the four groups. The authors also discuss how collocation and concordance analysis can be used to direct researchers to representative texts in order to carry out qualitative analysis.

Specialized software used:

WordSmith

www.lexically.net/wordsmith

Content Analysis

Content analysis adopts a quantitative, scientific approach to text analysis. Unlike CDA, content analysis is generally focused on texts themselves rather than texts' relations to their social and historical contexts. One of the classic definitions of content analysis defines it as "a research technique for the objective, systematic-quantitative description of the manifest content of communication" (Berelson, 1952, p. 18). At a practical

level, content analysis involves the development of a coding frame that is applied to textual data. It mainly consists of breaking down texts into pertinent units of information in order to permit subsequent coding and categorization.

Krippendorff's (2013) classic textbook *Content Analysis* is the standard reference for work in this area. Many of the research design principles and sampling techniques covered in Chapter 5 of this textbook are shared with content analysis, although Krippendorff's book goes into much greater detail on statistical sampling of texts and units of texts, as well as on statistical tests of interrater reliability.

Foucauldian Analysis

The philosopher and historian Foucault (1973) developed an influential conceptualization of intertextuality that differs significantly from Fairclough's conceptualization in CDA. Rather than identifying the influence of external discourses within a text, for Foucault the meaning of a text emerges in reference to discourses with which it engages in dialogue. These engagements may be explicit or, more often, implicit. In Foucauldian intertextual analysis, the analyst must ask each text about its presuppositions and with which discourses it dialogues. The meaning of a text therefore derives from its similarities and differences with respect to other texts and discourses and from implicit presuppositions within the text that can be recognized by historically informed close reading.

Foucauldian analysis of texts is performed in many theoretical and applied research fields. For instance, a number of studies have used Foucauldian intertextual analysis to analyze forestry policy (see Winkel, 2012, for an overview). Researchers working in Europe (e.g., Berglund, 2001; Franklin, 2002; Van Herzele, 2006), North America, and developing countries (e.g., Asher & Ojeda, 2009; Mathews, 2005) have used Foucauldian analysis to study policy discourses regarding forest management, forest fires, and corporate responsibility.

Another example of Foucauldian intertextual analysis is a sophisticated study of the professional identities of nurses by Bell, Campbell, and Goldberg (2015). Bell and colleagues argued that nurses' professional identities should be understood in relation to the identities of other occupational categories within the health care field. The authors collected their data from PubMed, a medical research database. Using PubMed's own user interface, the authors acquired the abstracts for research papers that used the terms *service* or *services* in the abstract or key words for a period from 1986 to 2013. The downloaded abstracts were added to an SQLite

database, which was used to generate comma-separated values (CSV) files with abstracts organized into 3-year periods. The authors then spent approximately 6 weeks of full-time work, manually checking the data for duplicates and other errors. The final sample included over 230,000 abstracts. Bell and colleagues then used the text analysis package Leximancer (see Appendix C) to calculate frequency and co-occurrence statistics for all concepts in the abstracts (see also Appendix F). Leximancer also produced concept maps (see Appendix G) to visually represent the relationships between concepts. The authors further cleaned their data after viewing these initial concept maps and finding a number of irrelevant terms and then used Leximancer to analyze the concept of nursing in terms of its co-occurrence with other concepts.

Analysis of Texts as Social Information

Another category of text analysis treats texts as reflections of the practical knowledge of their authors. This type of analysis is prevalent in grounded theory studies (see Chapter 4) as well as in applied studies of expert discourses. Interest in the informative analysis of texts is due in part to its practical value, because user-generated texts can potentially provide analysts with reliable information about social reality. Naturally, the quality of information about social reality that is contained in texts varies according to the level of knowledge of each individual who has participated in the creation of the text, and the information that subjects provide is partial insofar as it is filtered by their own particular point of view.

An example of analysis of texts as social information is a 2012 psychological study by Colley and Neal on the topic of organizational safety. Starting with small representative samples of upper managers, supervisors, and workers in an Australian freight and passenger rail company, Colley and Neal conducted open-ended interviews with members of the three groups. These were transcribed and analyzed using Leximancer (see Appendix C) for map analysis (see also Appendix G). Comparing the concept maps produced for the three groups revealed significant differences between the “safety climate schema” of upper managers, supervisors, and workers.

Challenges and Limitations of Using Online Data

Having introduced text mining and text analysis, in this section we review some lessons that have been

learned from other fields about how best to adapt social science research methods to data from online environments. This section is short but critically important for students who plan to perform research with data taken from social media platforms and websites.

Methodologies such as text mining that analyze data from digital environments offer potential cost- and time-efficiency advantages over older methods (Hewson & Laurent, 2012; Hewson, Yule, Laurent, & Vogel, 2003), as the Internet provides ready access to a potentially vast, geographically diverse participant pool. The speed and global reach of the Internet can facilitate cross-cultural research projects that would otherwise be prohibitively expensive. It also allows for the emergence of patterns of social interactions, which are elaborate in terms of their richness of communication exchange but where levels of anonymity and privacy can be high. The Internet's unique combination of digital archiving technologies and users' perceptions of anonymity and privacy may reduce social desirability effects (where research participants knowingly or unknowingly attempt to provide researchers with socially acceptable and desirable, rather than accurate, information). The unique attributes of Internet-based technologies may also reduce biases resulting from the perception of attributes such as race, ethnicity, and sex or gender, promoting greater candor. The convenience of these technologies can also empower research participants by allowing them to take part in study procedures that fit their schedules and can be performed within their own spaces such as at home or in a familiar work environment.

While Internet-based research has many advantages (see Hewson, Vogel, & Laurent, 2015), Internet-based data have a number of serious drawbacks for social science research. One major disadvantage is the potentially biased nature of Internet-accessed data samples. **Sample bias** is one of the most fundamental and difficult to manage challenges associated with Internet-mediated research (see Chapter 5). Second, as compared to offline methods, Internet-based data are often characterized by reduced levels of researcher control. This lack of control arises mainly from technical issues, such as users' different hardware and software configurations and network traffic performance. Research participants working with different hardware platforms, operating systems, and browsers may experience social media services and online surveys very differently, and it is often extremely difficult for researchers to fully appreciate differences in participants' experiences. In addition, hardware and software failures may lead to unpredicted effects, which may cause problems. Because of the lack of researcher presence, in Internet-based research there is often a lack of researcher control over and knowledge of variations in participants' behaviors and the participation context. This may cause problems related to the extent to which researchers can gauge participants' intentions and levels of sincerity and honesty during a study, as researchers lack nonverbal cues to evaluate participants compared with face-to-face communication.

Despite these weaknesses, scholars have long recognized digital technologies' potential as research tools. While social researchers have occasionally developed brand-new Internet-based methodologies, they have also adapted preexisting research methods for use with evolving digital technology. Because a number of broadly applicable lessons have been learned from these adaptation processes, in the remainder of this chapter we briefly review some of the most widely used social science research methods that have been adapted to Internet-related communication technologies and some of the lessons learned from each. We discuss offline and online approaches to *social surveys*, *ethnography*, and *archival research* but do not cover online focus groups (Krueger & Casey, 2014) or experiments (Birnbaum, 2000). While focus groups and experiments are both important and widely used research methods, we have found that the lessons learned from developing online versions of these methods are less applicable to text mining than lessons learned from the former three.

Social Surveys

Social surveys are one of the most commonly used methods in the social sciences, and researchers have been working with online versions of surveys since the 1990s. Traditional telephone and paper surveys tend to be costly, even when using relatively small samples, and the costs of a traditional large-scale survey using mailed questionnaires can be enormous. Although the costs of online survey creation software and web survey services vary widely, by eliminating the need for paper, postage, and data entry costs, online surveys are generally less expensive than their paper- and telephone-based equivalents (Couper, 2000; Ilieva, Baron, & Healey, 2002; Yun & Trumbo, 2000). Online surveys can also save researchers time by allowing them to quickly reach thousands of people despite possibly being separated by great geographic distances (Garton, Haythornthwaite, & Wellman, 2007). With an online survey, a researcher can quickly gain access to large populations by posting invitations to participate in the survey to newsgroups, chat rooms, and message boards. In addition to their cost and time savings and overall convenience, another advantage of online surveys is that they exploit the ability of the Internet to provide access to groups and individuals who would be difficult, if not impossible, to reach otherwise (Garton et al., 1997).

While online surveys have significant advantages over paper- and phone-based surveys, they bring with them new challenges in terms of applying traditional survey research methods to the study of online behavior. Online survey researchers often encounter problems regarding sampling, because relatively little may be known

about the characteristics of people in online communities aside from some basic demographic variables, and even this information may be questionable (Walejko, 2009). While attractive, features of online surveys themselves, such as multimedia, and of online survey services, such as use of company e-mail lists to generate samples, can affect the quality of the data they produce in a variety of ways.

The process of adapting social surveys to online environments offers a cautionary lesson for text mining researchers. The issue of user demographics casts a shadow over online survey research just as it does for text mining, because in online environments it is very difficult for researchers to make valid inferences about their populations of interest. The best practice for both methodologies is for researchers to carefully plan and then explain in precise detail their sampling strategies (see Chapter 5).

Ethnography

In the 1990s, researchers began to adapt ethnographic methods designed to study geographically situated communities to online environments which are characterized by relationships that are technologically mediated rather than immediate (Salmons, 2014). The result is **virtual ethnography** (Hine, 2000) or **netnography** (Kozinets, 2009), which is the ethnographic study of people interacting in a wide range of online environments. Kozinets, a netnography pioneer, argues that successful netnography requires researchers to acknowledge the unique characteristics of these environments and to effect a “radical shift” from offline ethnography, which observes people, to a mode of analysis that involves recontextualizing conversational acts (Kozinets, 2002, p. 64). Because netnography provides more limited access to fixed demographic markers than does ethnography, the identities of discussants are much more difficult to discern. Yet netnographers must learn as much as possible about the forums, groups, and individuals they seek to understand. Unlike in traditional ethnographies, in the identification of relevant communities, online search engines have proven invaluable to the task of learning about research populations (Kozinets, 2002, p. 63).

Just as the quality of social survey research depends on sampling, netnography requires careful case selection (see Chapter 5). Netnographers must begin with specific research questions and then identify online forums appropriate to these questions (Kozinets, 2009, p. 89).

Netnography’s lessons for text mining and analysis are straightforward. Leading researchers have shown that

for netnography to be successful, researchers must acknowledge the unique characteristics of online environments, recognize the importance of developing and explaining their data selection strategy, and learn as much as they possibly can about their populations of interest. All three lessons apply to text mining research that analyzes user-generated data mined from online sources.

Historical Research Methods

Archival research methods are among the oldest methods in the social sciences. The founding fathers of sociology—Marx, Weber, and Durkheim—all did historical scholarship based on archival research, and today, archival research methods are widely used by historians, political scientists, and sociologists.

Historical researchers have adapted digital technology to archival research in two waves. The first occurred in the 1950s and 1960s when, in the early years of accessible computers, historians taught themselves statistical methods and programming languages. Adopting quantitative methods developed in sociology and political science, during this period historians made lasting contributions in the areas of "social mobility, political identification, family formation, patterns of crime, economic growth, and the consequences of ethnic identity" (Ayers, 1999). Unfortunately, however, that quantitative social science history collapsed suddenly, the victim of its own inflated claims, limited method and machinery, and changing academic fashion. By the mid-80s, history, along with many of the humanities and social sciences, had taken the linguistic turn. Rather than SPSS guides and codebooks, innovative historians carried books of French philosophy and German literary interpretation. The social science of choice shifted from sociology to anthropology; texts replaced tables. A new generation defined itself in opposition to social scientific methods just as energetically as an earlier generation had seen in those methods the best means of writing a truly democratic history. The first computer revolution largely failed (Ayers, 1999).

Beginning in the 1980s, historians and historically minded social scientists began to reengage with digital technologies. While today historical researchers use digital technologies at every stage of the research process, from professional communication to multimedia presentations, **digital archives** have had perhaps the most profound influence on the practice of historical research. Universities, research institutes, and private companies have digitized and created accessible archives of massive volumes of historical documents. Histo-

rians recognize that these archives offer tremendous advantages in terms of the capacity, flexibility, accessibility, flexibility, diversity, manipulability, and interactivity of research (Cohen & Rosenzweig, 2005). However, digital research archives also pose dangers in terms of the quality, durability, and readability of stored data. There is also a potential for inaccessibility and monopoly and also for digital archives to encourage researcher passivity (Cohen & Rosenzweig, 2005).

There are lessons to be learned from digital history for text mining and text analysis, particularly from the sudden collapse of the digital history movement of the 1950s and 1960s. In light of the failure of that movement, it is imperative that social scientists working with text mining tools recognize the limitations of their chosen methods and not make imperious or inflated claims about these tools' revolutionary potential. Like all social science methods, text mining methods have benefits and drawbacks that must be recognized from the start and given consideration in every phase of the research process. And text mining researchers should be aware of historians' concerns about the quality of data stored in digital archives and the possibility for digital archives to encourage researcher passivity in the data gathering phase of research.

Conclusion

This chapter has introduced text mining and text analysis methodologies, provided an overview of the major approaches to text analysis, and discussed some of the risks associated with analyzing data from online sources. Despite these risks, social and computer scientists are developing new text mining and text analysis tools to address a broad spectrum of applied and theoretical research questions, in academia as well as in the private and public sectors.

In the chapters that follow, you will learn how to find data online (Chapters 2 and 6), and you will learn about some of the ethical (Chapter 3) and philosophical and logical (Chapter 4) dimensions of text mining research. In Chapter 5, you will learn how to design your own social science research project. Parts II, IV, and V review specific text mining techniques for collecting and analyzing data, and Chapter 17 in Part VI provides guidance for writing and reporting your own research.

Key Terms (see Glossary)

Highlights

- Text mining processes include methods for acquiring digital texts and analyzing them with NLP and advanced statistical methods.
- Text mining is used in many academic and applied fields to analyze and predict public opinion and collective behavior.
- Text analysis began with analysis of religious texts in the Middle Ages and was developed by social scientists starting in the early 20th century.
- Text analysis in the social sciences involves analyzing transcribed interviews, newspapers, historical

and legal documents, and online data.

- Major approaches to text analysis include analysis of discourse positions, conversation analysis, CDA, content analysis, intertextual analysis, and analysis of texts as social information.
- Advantages of Internet-based data and social science research methods include their low cost, unobtrusiveness, and use of unprompted data from research participants.
- Risks and limitations of Internet-based data and research methods include limited researcher control, possible sample bias, and the risk of researcher passivity in data collection.

Review Questions

- What are the differences between text mining and text analysis methodologies?
- What are the main research processes involved in text mining?
- How is analysis of discourse positions different from conversation analysis?
- What kinds of software can be used for analysis of discourse positions and conversation analysis?

Discussion Questions

- If you were interested in conducting a CDA of a contemporary discourse, what discourse would you study? Where would you find data for your analysis?
- How do researchers choose between collecting data from offline sources, such as in-person interviews, and online sources, such as social media platforms?
- What are the most critical problems with using data from online sources?
- If you already have an idea for a research project, what are likely to be the most critical advantages and disadvantages of using online data for your project?
- What are some ways text mining research be used to benefit science and society?

Developing a Research Proposal

Select a social issue that interests you. How might you analyze how people talk about this issue? Are there

differences between people from different communities and backgrounds in terms of how they think about this issue? Where (e.g., offline, online) do people talk about this issue, and how could you collect data from them?

Further Reading

Ayers, E. L. (1999). *The pasts and futures of digital history*. Retrieved June 17, 2015, from http://www.vcdh.virginia.edu/PastsFutures.html

Bauer, M. W., Bicquelet, A., & Suerdem, A. K. (Eds.), *Textual analysis. SAGE benchmarks in social research methods* (Vol. 1). Thousand Oaks, CA: Sage.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Kuckartz, U. (2014). *Qualitative text analysis: A guide to methods, practice, and using software.* Thousand Oaks, CA: Sage.

Roberts, C. W. (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts.* Mahwah, NJ: Lawrence Erlbaum.

https://doi.org/10.4135/9781506336985

Appendix A Data Sources for Text Mining

The American Presidency Project

www.presidency.ucsb.edu

The American Presidency Project is one of the most comprehensive collections of web resources on the American presidency, including documents, public papers, executive orders, addresses, press conferences, debates, election data, and approval ratings data. The American Presidency Project was established in 1999 as a collaboration between John T. Woolley and Gerhard Peters at the University of California, Santa Barbara. The archives contain 116,994 documents related to the study of the presidency.

arXiv Bulk Data Access

https://arxiv.org/help/bulk_data

This is a continuously updated list of high-quality open data sets in public domains.

Category:Dataset

http://wiki.urbanhogfarm.com/index.php/Category:Dataset

A community effort, Category:Dataset seeks to aggregate public data sets related to social media and online communities.

CMU Movie Summary Corpus

www.cs.cmu.edu/~ark/personas

This page provides links to a data set of movie plot summaries and associated metadata collected by David Bamman, Brendan O'Connor, and Noah A. Smith at the Language Technologies Institute and Machine Learning Department at Carnegie Mellon University.

Congressional and Federal Government Web Harvests

https://webharvest.gov

Since 2006, the U.S. National Archives and Records Administration has harvested Congressional websites at the end of each Congress. They also did a harvest of all federal websites for the 2004 presidential transition.

Congressional Record

https://www.gpo.gov/fdsys/browse/collection.action?collectionCode=CREC

The *Congressional Record* is the official record of the proceedings and debates of the U.S. Congress. The *Congressional Record* began publication in 1873 and is still published today. It is published daily when Congress is in session, and its documents are available as ASCII text and in Adobe PDF.

Consumer Complaint Database

https://catalog.data.gov/dataset/consumer-complaint-database

From the Consumer Financial Protection Bureau, these are complaints received about financial products and services available as comma-separated values (CSV) files and in other formats as well.

Corpus of Contemporary American English

http://corpus.byu.edu/coca

The Corpus of Contemporary American English (COCA) is the largest public access corpus of English and the only large and balanced corpus of American English. The corpus contains more than 520 million words of text and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts.

DocumentCloud

https://www.documentcloud.org

DocumentCloud runs every document you upload through Thomson Reuters OpenCalais, which tags the people, places, companies, facts, and events in the document.

EBSCO Newspaper Source

https://www.ebscohost.com/public/newspaper-source

This database provides full text for over 400 national (United States), international, and regional newspapers. It also offers television and radio news transcripts from major networks.

GloWbE: Corpus of Global Web-Based English

http://corpus.byu.edu/glowbe

The Corpus of Global Web-Based English (GloWbE), created by Mark Davies of Brigham Young University, is composed of 1.9 billion words from 1.8 million webpages in 20 different English-speaking countries. The corpus was released in 2013 and is related to other large corpora including the 520 million-word Corpus of Contemporary American English (COCA) and the 400 million-word Corpus of Historical American English (COHA). Together, these three corpora allow researchers to examine variation in English by dialect, genre, and over time. Data in GloWbE comes in three formats including tables for relational databases, word/lemma/part of speech, or text.

HathiTrust

https://www.hathitrust.org/datasets

HathiTrust is a partnership of major research institutions and libraries working to ensure that the cultural record is preserved and accessible long into the future. HathiTrust's documents include non-Google-digitized volumes, which are freely available, and Google-digitized volumes, which are available through an agreement with Google. Within each category, there is a distinction between public domain works available only in the United States and public domain works available anywhere in the world. The non-Google-digitized volumes include approximately 550,000 public domain volumes as of March 2015, which are primarily English-language materials published prior to 1923. The Google-digitized volumes include approximately 4.8 million public domain volumes as of March 2015, representing a wide variety of languages, subjects, and dates.

Internet Archive

https://archive.org

Founded in 1996 and located in San Francisco, the Internet Archive is a nonprofit organization that was founded to build an Internet library. Its purposes include offering permanent access for researchers, historians, scholars, people with disabilities, and the general public to historical collections that exist in digital format. Internet Archive contains millions of free books, movies, software, and music.

JSTOR for Research

http://dfr.jstor.org

JSTOR is a popular digital library of academic journals, books, and primary sources. JSTOR's Data for Research service is a free service for researchers wishing to analyze content on JSTOR through a variety of lenses and perspectives.

LexisNexis Academic

https://www.lexisnexis.com/hottopics/lnacademic

LexisNexis Group is a corporation providing computer-assisted legal research as well as business research and risk management services. Their news database includes news from over 10,000 sources.

Observatory on Social Media

https://osome.iuni.iu.edu

The Observatory on Social Media (also known as Truthy) is an informal nickname associated with a research project of the Center for Complex Networks and Systems Research at the Indiana University School of Informatics and Computing. The project aims to study how information spreads on social media, such as Twitter. The project has focused on domains such as news, politics, social movements, scientific results, and trending social media topics. Researchers develop theoretical computer models and validate them by analyzing public data, mainly from the Twitter streaming application programming interface (API). Social media posts available through public APIs are processed without human intervention or judgment to visualize and study the spread of millions of memes. An important goal of the project is to help mitigate misuse and abuse of social media by helping us better understand how social media can be potentially abused.

OpenLibrary

https://openlibrary.org/data

Open Library is an open, editable library catalog building toward a webpage for every book ever published.

Public.Resource.Org

https://public.resource.org

Public.Resource.Org contains bulk downloadable content harvested from government websites and other sources.

PubMed

https://www.ncbi.nlm.nih.gov/pubmed

PubMed includes more than 25 million citations for literature from biomedical fields. Some citations include links to full-text content.

Robots Reading *Vogue*

http://dh.library.yale.edu/projects/vogue

This project from Yale University is based on the ProQuest *Vogue* Archive (www.proquest.com/products-services/vogue_archive.html). *Vogue* is an American lifestyle magazine that has been continuously published for over a century. The archive contains over 2,700 covers, 400,000 pages, and six TB of data.

Text Creation Partnership

www.textcreationpartnership.org

The primary goal of the Text Creation Partnership is to create standardized, accurate XML/SGML encoded electronic text editions of early printed books. The partnership transcribes and encodes the page images of books from ProQuest's Early English Books Online, Gale Cengage's Eighteenth Century Collections Online, and Readex's Evans Early American Imprints. The resulting text files are jointly funded and owned by more than 150 libraries worldwide.

the @unitedstates project

https://theunitedstates.io

The @unitedstates project is a shared commons of data and tools for the United States that features work from people with the Sunlight Foundation, GovTrack.us, the *New York Times*, and the Electronic Frontier Foundation.

University of Oxford Text Archive

https://ota.ox.ac.uk

The University of Oxford Text Archive develops, collects, catalogs, and preserves electronic literary and linguistic resources for use in research, teaching, and learning. The Oxford Text Archive also gives advice on the creation and use of these resources and is involved in the development of standards and infrastructure for electronic language resources.

Yahoo Webscope Program

https://webscope.sandbox.yahoo.com/#datasets

The Yahoo Webscope Program data sets are a reference library of interesting and scientifically useful data sets for noncommercial use by academics and other scientists.

Appendix B Text Preparation and Cleaning Software

Text data are not always in a format that can be used for social science research. In many cases, you will need to clean your data, eliminating nonwords such as URLs, advertisements, and copied text (such as in e-mail chains). In some text mining projects, you will need to eliminate stop words (see Chapter 8) such as *and* and *the*. While there is no one standard list of English stop words, an Internet search will yield several such lists that you can use.

There are several tools and methods that can be used for cleaning texts, including simple Find and Replace commands in word processors and spreadsheets, regular expressions (regexes), and software.

Find and Replace

The most basic tool for cleaning texts is the Find and Replace text editor command in word processors such as Microsoft Word and Google Docs. To use Find and Replace effectively, you must know your data well, keeping your eyes on patterns and repetitions in your files. Before trying Replace All, start by replacing unwanted text one at a time until you are certain you are not replacing important content. Work around the parts that don't repeat, and use the existing structure of the data to your advantage.

Regexes

When there are patterns in a text file but not exact character matches, you can use regular expressions, or regexes. Microsoft Word, Google Docs, and Google Sheets all have regex functionality, which can be a tremendous time saver when working with large files.

In Microsoft Word, regex Find and Replace commands use wildcard characters, which are keyboard characters that can represent one or many characters. For instance, the asterisk (*) typically represents one or more characters, and the question mark (?) typically represents a single character. Regexes are combinations of literal and wildcard characters that you use to find and replace patterns of text. The literal text characters indicate text that must exist in the target string of text, and the wildcard characters indicate the text that can vary in the target string.

https://support.office.com/en-us/article/Find-and-replace-text-by-using-regular-expressions-Advanced-eeaa03b0-e9f3-4921-b1e8-85b0ad1c427f

In Google Docs, the command REGEXREPLACE allows you to replace part of a text string with a different text string using regular expressions. If you are using Google Sheets, you can use REGEXMATCH. You can learn more about Google's regular expressions here:

https://support.google.com/docs/answer/3098244

Regexes require some memorization and practice, but if you need to clean large documents and document collections they are a powerful and inexpensive option.

Software

While Find and Replace and regular expressions are useful tools for cleaning data, there are many software packages available that can help you clean, organize, and manage your document collections.

Adobe Acrobat

https://acrobat.adobe.com/us/en/acrobat.html

The full version of Adobe Acrobat (not Adobe Reader) allows you to convert PDF files into plain text quickly as well as perform top of the line optical character recognition (OCR). If you are having trouble with texts with formatting issues, you may consider trying Acrobat.

BBEdit

http://www.barebones.com/products/bbedit

BBEdit is a professional HTML and text editor for Mac designed for web authors and software developers. BBEdit's features include regular expression (regex) pattern matching, search and replace across multiple files, project definition tools, function navigation and syntax coloring for numerous source code languages,

code folding, FTP and SFTP open and save, AppleScript, Mac OS X Unix scripting support, text and code completion, and a complete set of HTML markup tools.

OpenRefine

http://openrefine.org

OpenRefine (formerly Google Refine) is a free, open source tool for cleaning data and transforming it from one format into another.

TextCleanr

www.textcleanr.com

TextCleanr is a simple-to-use web-based tool for fixing and cleaning up text when copying and pasting between applications. It is able to remove e-mail indents, find and replace, and clean up spacing and line breaks.

TextPipe

www.datamystic.com/textpipe

This software suite for text processing makes it easy to write filters to strip documents of HTML tags or other similar formatting tasks.

TextSoap

https://www.unmarked.com/textsoap

TextSoap automatically removes unwanted characters and can fix messed up carriage returns. It features over 100 built-in cleaners and has regular expression support.

Trifacta Wrangler

https://www.trifacta.com/products/wrangler

Originally known as Data Wrangler, Trifacta Wrangler is a text cleaning and formatting tool that can automatically find patterns in your data based on things you select and can even make suggestions as to what to do with those patterns. It also learns over time, so it is constantly improving the suggestion system.

UltraEdit

http://ultraedit.com

UltraEdit is a powerful Windows-based tool that can load and work with extremely large text files.

Appendix C General Text Analysis Software

For students who prefer to use commercial or free software rather than programming environments like Python or R, in this appendix we provide an overview of general text mining and text analysis software that has been used in social science research. The software packages include Leximancer, Linguistic Inquiry and Word Count (LIWC), RapidMiner, TextAnalyst, and WordStat.

Leximancer

https://info.leximancer.com

Leximancer, originally created at the University of Queensland in Australia, includes concept mapping and sentiment analysis tools. In Leximancer, a text block is the unit of analysis, and the software is Bayesian-based in that it "learns" from an uploaded data set that it reads iteratively. For concept mapping, it creates a network of concepts defined in text blocks of about a paragraph in size. For sentiment analysis, Leximancer maps the frequency and co-occurrence of concepts with a built-in thesaurus of sentiment terms (positive versus negative).

Leximancer has been used by the health researchers Bell, Campbell, and Goldberg (2015) in their Foucauldian analysis of nurses' professional identities and by the psychologists Colley and Neal (2012) in their study of organizational safety.

Linguistic Inquiry and Word Count

http://liwc.wpengine.com

Based on psychological research by James Pennebaker, Linguistic Inquiry and Word Count (LIWC) is a text analysis program that counts words based on psychological categories. It has been used in numerous studies of attentional focus, emotionality, social relationships, thinking styles, and personality differences (see Tausczik & Pennebaker, 2010). It has also been used in computer science studies (e.g., Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012).

RapidMiner

http://rapidminer.com

RapidMiner is an open source system for data mining available as a stand-alone application for data analysis and as a data mining engine. It is used for business and commercial applications as well as for research and education and supports all steps of the data mining process including data preparation, validation, and results visualization.

TextAnalyst

http://megaputer.com/site/textanalyst.php

TextAnalyst maps out semantic relationships between specific terms within a document or set of documents to highlight thematic structures within the text. It quantifies terms based on their overall relevance to the text as a whole as well as their relationships to each other, generating a semantic network of the interrelated themes within a text document through the application of linguistic rules and an “artificial neural network” program that approximates human cognition. TextAnalyst has been used by the sociologists Adams (2009), Roscigno (Adams & Roscigno 2005), and Ignatow (2009).

WordStat

https://provalisresearch.com/products/content-analysis-software

WordStat works with QDA Miner (see Appendix D) and offers keyword-in-context, keyword retrieval, dictionary building, machine learning, and visualization capabilities.

Research in the Spotlight

Using TextAnalyst to Study Collective Identity

Adams, J. (2009). Bodies of change: A comparative analysis of media representations of body modification practices. *Sociological Perspectives*, *52*(1), 103–129.

The sociologist Adams examined how mainstream media represent cosmetic surgery, tattooing, and body piercing by analyzing 72 newspaper articles using TextAnalyst. Adams found that cosmetic surgery and tattooing are positively presented as consumer lifestyle options, while piercing is often negatively framed as an unhealthy and problematic practice. The risks associated with cosmetic surgery and tattooing are frequently downplayed, as are tattooing's associations with deviance, while the potential risks related to body piercing are overemphasized. Gender is also a prominent framing device, often used to reinforce normative appearance expectations.

Adams, J., & Roscigno, V. (2005). White supremacists, oppositional culture and the world wide web. *Social Forces*, *84*(2), 759–778.

As in the 2009 study of body modification by Josh Adams, in this study from 2005 Adams and the sociologist Roscigno used TextAnalyst to investigate how white supremacist organizations use the Internet. Specifically, Adams and Roscigno investigated how these groups recruit members, build collective identities, and organize their activities. The authors used TextAnalyst to construct semantic network graphs (see Appendix G on concept maps) of thematic content from major white supremacist websites and to delineate patterns and thematic associations relative to three aspects of social movement culture: identity, interpretational framing of cause and effect, and political efficacy. They found that nationalism, religion, and definitions of responsible citizenship are interwoven with race to create a sense of collective identity for members of these groups as well as for potential recruits. These groups use interpretative frameworks that simultaneously identify threatening social issues and provide corresponding recommendations for social action. Adams and Roscigno discussed how the Internet has been integrated into white supremacist groups' tactical repertoires.

Appendix D Qualitative Data Analysis Software

Narrative analysis (see Chapter 10), thematic analysis (see Chapter 11), metaphor analysis (see Chapter 12), and other forms of qualitative and mixed method text analysis can be performed without the help of highly specialized software (beyond word processors and spreadsheets). But specialized software can help to expand the scope, methodological sophistication, and rigor of text analysis research. The most popular software used for text analysis is known as computer-assisted qualitative data analysis software (CAQDAS, or qualitative data analysis software [QDAS] for short). QDAS packages are tools for organizing collections of documents so that they can be more efficiently and effectively analyzed qualitatively, although as we will see several QDAS packages include modules for statistical analysis and data visualization. Such software is widely used in psychology, sociology, and marketing research, and typically includes tools for content searching, coding or labeling text, linking text units, querying, writing and annotation, and visualizing results as maps, networks, or word clouds (see Appendix G).

Versions of QDAS have been around since the 1980s and have been used to assist content analysis, discourse analysis, grounded theory analysis, and mixed method projects. The first version of the QDAS program NUD*IST was released in 1981, and ATLAS and WinMAX were released in 1989. These software packages subsequently evolved into more developed forms: WinMAX into MAXQDA, ATLAS into ATLAS.ti, and NUD*IST into NVivo.

QDAS packages perform several interrelated functions for researchers. First and foremost they allow researchers to code and retrieve samples of text. They also allow researchers to use coded text to build theoretical models of the social, psychological, cognitive, and linguistic processes that are thought to have generated the text. Their interfaces also allow for relatively easy text retrieval and for management of and navigation within large document collections. In addition to these core functions, as we will see, many software packages allow for visualization and statistical analysis of the interrelationships between coded textual units.

A central feature of QDAS is the ability to set up rules to apply labels to texts. QDAS packages offer a variety of text coding techniques that allow for code and retrieve functionality, including *in vivo coding*, an inductive method where a word or short phrase taken from the text itself is the code or label (King, 2008). Other forms of coding include *free coding*, which involves assigning any code to arbitrary sequences of data; *contextual coding*, in which users label text in such a way as to allow them to quickly navigate to view the labeled text in context; *automatic coding*, which involves assigning codes automatically to search results; and even

artificial intelligence-based *software-generated coding*, in which the software suggests codes based on its own analysis of the text.

QDAS packages feature a number of different types of text search tools, including simple searches; Boolean searches using the Boolean operators AND, OR, and NOT; placeholder searches that allow you to use placeholders for certain characters; and proximity searches that allow you to retrieve combinations of two or more text strings and/or codes that occur in a definable proximity to each other. Fuzzy searches, or "approximation searches," are as of this writing exclusive to NVivo. These allow you to perform searches that retrieve textual data even if the data contain typographic errors. Combination searches involve combinations of some of the previously mentioned types of searches.

In addition to coding and searching texts, QDAS packages provide a variety of different tools for annotation (e.g., memo writing and storage) and for producing output in different formats, from variable diagrams and network diagrams for visualizing theoretical models, to word clouds. Most software allows users to export data on code and word frequencies to allows for statistical analysis with appropriate statistical packages such as SPSS or STATA, or else include statistical tools for analyzing word frequencies, cross-tabulations, clusters, and word co-occurrence matrices.

Although you may be tempted to use the software that is most familiar or available to you, it is worth investing some time in carefully calculating the benefits and disadvantages of the various software tools that might be used for your project. Because the learning curve for some of these packages is steep and the time commitment involved substantial, it is important to choose the best tool for the job.

Different types of research projects require QDAS packages with different sets of features. Some qualitative data analysis software packages feature dashboards that are especially easy to use, others have more powerful project management and data organization tools, while still others allow users to easily explore and interact with their data.

While QDAS packages are popular and widely used in several disciplines, the value of using software in qualitative analysis has been vigorously debated. Coffey, Holbrook, and Atkinson (1996), Macmillan (2005), and Goble, Austin, Larsen, Kreitzer, and Brintnell (2012) are good places to start for critical appraisals of QDAS (see the Further Reading section).

Loughborough University's QDAS site provides some useful guidelines for matching software to research project requirements. They recommend MaxQDA or QDA Miner for mixed methods projects; NVivo and ATLAS.ti for discourse analysis (see Chapter 1); and ATLAS.ti, HyperRESEARCH, or Qualrus for virtual

ethnography (see Chapter 1). It is, of course, important to refer to the software websites to keep up to date on new features, as new versions of both commercial and open source software are released frequently.

Commercial Software

ATLAS.ti

http://atlasti.com

One of the first and most highly developed QDAS tools, it allows coded data to be exported for analysis with statistical packages such as SPSS.

Dedoose

www.dedoose.com

A web-based qualitative and mixed methods research application, Dedoose builds on tools available in its predecessor, EthnoNotes. Dedoose is specifically designed to support the concurrent analysis of large amounts of mixed data by teams of geographically dispersed researchers.

f4analyse

http://www.audiotranskription.de/english/f4-analyse

This is a basic, competitively priced QDAS tool from Germany that is easy to use.

HyperRESEARCH

http://www.researchware.com/products/hyperresearch.html

HyperRESEARCH is a QDAS for the Mac OS featuring advanced multimedia capabilities.

Kwalitan

http://www.kwalitan.nl

Designed to assist in the development of grounded theories, this software from the Netherlands enables hierarchical coding and the navigation of data with Boolean searches.

MAXQDA

http://maxqda.com

MAXQDA is a sophisticated package with statistical and visualization add-ons available.

NVivo

http://www.qsrinternational.com/product

NVivo features relatively elaborate organizing functions that allow users to link together text data in a variety of ways.

QDA Miner

https://provalisresearch.com/products/qualitative-data-analysis-software

A sophisticated QDAS tool that integrates with SimStat, a statistical data analysis module, and WordStat, a quantitative content analysis and text mining module.

Qualrus

http://qualrus.com

Qualrus is a QDAS tool that is “portable” for use on multiple platforms (Mac, Windows).

Quirkos

https://www.quirkos.com

Quirkos is an easy-to-use and competitively priced QDAS tool from the University of Edinburgh.

Although free trial versions of most commercial QDAS packages are available, the full versions can be expensive, particularly for single users who do not have access to a group license. So you may want to explore some of the many free and open source QDAS tools that may meet their needs. If you already use the programming language R, or are considering using it for quantitative analysis, you can use the RQDA package to combine text coding with the statistical power of R. RQDA is probably the most advanced of all the free QDAS packages. It allows users to perform word cloud analysis (see Appendix G), create queries for complex cross-coding retrieval, program auto-coding commands, plot the relationship between codes, and export data as spreadsheets. RQDA also features a very intuitive user interface.

Free and Open Source Qualitative Data Analysis Software

There are dozens of free and open source QDAS/CAQDAS tools available, including QDA Miner Lite, which is a free version of QDA Miner with limited features for both PC and Mac, Open Code (PC only), Saturate (cloud), and Coding Analysis Toolkit (CAT; cloud). Some of these packages are quite sophisticated: Text Analysis Markup System (TAMS), and RQDA allow both inductive and deductive coding as well as coding memos, can support hierarchical or structured coding, provide basic coding statistics, and perform text and coding retrieval. For quick and simple coding, Open Code and Saturate are easy to use, although they only allow one code per predefined text segment. Saturate is particularly well suited for coding with more than one analyst.

AQUAD

www.aquad.de/en

Aquad is a German open source QDAS package with sophisticated features including Boolean search.

Cassandre

www.cassandre.ulg.ac.be

Cassandre is a free QDAS package for Windows, Mac, and Linux from Belgium. Most documentation is in French.

Coding Analysis Toolkit

http://cat.textifer.com

Coding Analysis Toolkit (CAT) is a free web-based tool from the University of Pittsburgh. CAT was designed to use mainly keystrokes rather than the mouse for coding and can import an ATLAS.ti project for quantitative analysis, though it has a coding mechanism built into itself as well.

CATMA

http://www.catma.de

CATMA is software for Windows, Mac OS, and Linux developed at the University of Hamburg mainly for humanities researchers.

Compendium

http://compendium.open.ac.uk/institute

Compendium is a general purpose sharing and collaboration tool from the Open University in the United Kingdom.

FreeQDA

http://freeqda.sourceforge.net

FreeQDA is an open source QDAS tool.

libreQDA

http://www.libreqda.edu.uy

This is a free Spanish-language QDAS tool developed in Uruguay.

Open Code

http://www.phmed.umu.se/english/units/epidemiology/research/open-code

Free from Umea University in Sweden, Open Code was originally developed for use with grounded theory but now is a general-purpose qualitative data analysis tool.

QDA Miner Lite

https://provalisresearch.com/products/qualitative-data-analysis-software/freeware

QDA Miner Lite is an easy-to-use version of QDA Miner that offers basic QDAS features.

RQDA

http://rqda.r-forge.r-project.org

RQDA is a package for the popular programming language R. Used with R, it performs both qualitative and quantitative analysis.

Saturate

http://onlineqda.hud.ac.uk/Step_by_step_software/Saturate

Saturate is an online QDAS tool from the University of Huddersfield in the United Kingdom.

Text Analysis Markup System

https://sourceforge.net/projects/tamsys

Text Analysis Markup System (TAMS) is an open source QDAS tool.

Text Analysis Markup System Analyzer

http://tamsys.sourceforge.net

Text Analysis Markup System (TAMS) Analyzer works with TAMS to allow users to efficiently assign codes to text passages.

QDAS Tips

- Different packages are best suited to different types of research projects. Don't rush when selecting a package.
- Check that the package you select can output codes in formats you can use at later stages of your research—for example, for statistical analysis or visualization.

Internet Resources

CAQDAS Networking Project

https://www.surrey.ac.uk/sociology/research/researchcentres/caqdas

Loughborough University's CAQDAS Site

www.restore.ac.uk/lboro/research/software/caqdas_comparison.php

Further Reading

Coffey, A., Holbrook, B., & Atkinson, P. (1996). Qualitative data analysis: Technologies and representations. *Sociological Research Online*, *1*(1). Retrieved from http://www.socresonline.org.uk/1/1/4.html

Goble, E., Austin, W., Larsen, D., Kreitzer, L. E., & Brintnell, S. (2012). Habits of mind and the split-mind effect: When computer-assisted qualitative data analysis software is used in phenomenological research. *Forum: Qualitative Social Research*, *13*(2). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1709

Macmillan, K. (2005). More than just coding? Evaluating CAQDAS in a discourse analysis of news texts. *Forum: Qualitative Social Research*, *6*(3). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/28/59

Appendix E Opinion Mining Software

Opinion mining (sentiment analysis) can be performed with Python or in other programming environments, but there are also many opinion mining software packages available. While these are designed mainly for business intelligence, a few have been used for social science research.

Lexicoder

www.lexicoder.com
Lexicoder is a Java-based, multiplatform package for automated content analysis of text that is freely available for academic use. Lexicoder features a sentiment dictionary designed to capture sentiment in political texts.

OpinionFinder

http://mpqa.cs.pitt.edu/opinionfinder
OpinionFinder is a system that processes documents and automatically identifies subjective sentences as well as various aspects of subjectivity within sentences, including agents who are sources of opinion, direct subjective expressions and speech events, and sentiment expressions. OpinionFinder was developed by researchers at the University of Pittsburgh, Cornell University, and the University of Utah.

RapidMiner Sentiment Analysis

https://rapidminer.com/solutions/sentiment-analysis
RapidMiner is an analytics platform for web crawling and mining (see Appendix C) that provides an integrated environment for machine learning, data mining, text mining, predictive analytics, and business analytics.

SAS Sentiment Analysis Studio

https://www.sas.com/en_us/software/analytics/sentiment-analysis.html

SAS is widely used in business but is not widely used in the social sciences due to its high cost.

Appendix F Concordance and Keyword Frequency Software

As is discussed briefly in Chapter 1, concordancing arose out of a practical need for biblical scholars to be able to alphabetize and cite words and passages in the bible. Linguists began using computers to create concordances in the 1950s, and literary scholars as well as library and information scientists began working with computer-generated concordances that analyzed keywords in context (KWIC) in the 1970s. The term *corpus linguistics* did not come into common usage until the early 1980s, and social scientists did not begin to use corpus linguistics tools until the 1990s, when Fairclough and other critical discourse analysis (CDA) researchers (see Chapter 1) began to experiment with them.

Adelaide Text Analysis Tool

https://www.adelaide.edu.au/carst/resources-tools/adtat

The Adelaide Text Analysis Tool (AdTAT) is a cross-platform tool that can conduct basic word and phrase searches and searches for associated words and phrases. It provides frequency lists of words appearing both left and right of search terms, can print and save results, and can assist in constructing corpora.

AntConc

www.laurenceanthony.net/software/antconc

AntConc is a free corpus analysis toolkit for concordancing and text analysis. It works on multiple platforms and includes a concordancer, word and keyword frequency generators, tools for cluster and lexical bundle analysis, and a word distribution plot. AntConc also offers the choice of simple wildcard searches or regular expression (regex) searches (see Appendix B) and features an intuitive user interface.

Simple Concordance Program

www.textworld.com/scp

Simple Concordance Program is a free concordance and word-listing program for Windows and

Mac that lets you create word lists and search natural language text files for words, phrases, and patterns. It is able to read texts written in English, French, German, Polish, Greek, Russian, and other languages. In her 2014 book on social movements and social class, the sociologist Leondar-Wright used Simple Concordance Program to analyze class speech differences.

TextSTAT

http://neon.niederlandistik.fu-berlin.de/en/textstat

TextSTAT is a simple program that reads plain text files and HTML files directly from the Internet and produces word frequency lists and concordances from these files. It allows you to use regular expressions (regexes; see Appendix B) and can cope with many different languages and file encodings. Social science studies using TextSTAT include a 2010 study by the communications researchers Hellsten, Dawson, and Leydesdorff that used semantic maps (see Appendix G) to analyze newspaper debates on artificial sweeteners published in the *New York Times* between 1980 and 2006.

Wmatrix

http://ucrel.lancs.ac.uk/wmatrix

Wmatrix is a web-based software package that features corpus annotation tools and standard corpus linguistic methodologies such as frequency lists and concordances.

WordSmith

www.lexically.net/wordsmith

WordSmith is a popular package for Windows that offers tools for analyzing keywords in context (KWIC), analyzing word co-occurrences, and building dictionaries. Published by Lexical Analysis Software and Oxford University Press since 1996, it has been used by CDA (see Chapter 1) researchers including Fairclough (2006). It has also been used by media researchers (e.g., Ensslin & Johnson, 2006).

Appendix G Visualization Software

Software tools for visualization of patterns of word use and themes in texts are increasingly popular in the social sciences. In this appendix, we survey visualization tools that can be used with qualitative data analysis software (QDAS) packages (see Appendix D) as well as in combination with other types of software.

While the field of visualization is developing rapidly and exciting new tools for visualizing patterns in texts are introduced regularly, you should recognize that these tools have some limitations. The majority of visualization techniques ultimately transform qualitative data into quantifiable segments, an approach to analysis that may be antithetical to the goals of qualitative research methods (see Biernacki, 2014). If done poorly, visualizations may distract from the meaning and power of your analysis. Visual transformation of texts may result in a loss of emotional tone and nuances of meaning and may also lead to the impression that an analysis is less ambiguous and contradictory than it actually is. Thus, in addition to visualizations, you should consider including text excerpts or longer narratives to explore your texts and communicate your findings to your readers.

There are many software tools available to visually represent words and themes in texts (see Henderson & Segal, 2013), including correspondence analysis (LeRoux & Rouanet, 2010), path and network diagrams (Durland & Fredericks, 2005), decision trees (Ryan & Bernard, 2010), and tools for visualization of sentiment analysis (Gregory et al., 2006). In this appendix, we survey several of the most accessible visualization tools for text mining and text analysis including word clouds, word trees and phrase nets, and matrices and maps.

Word Clouds

Word clouds provide a visual display of word counts from one or more texts. The more frequently a word appears, the larger the word is displayed in a word cloud visual (Viégas & Wattenberg, 2008). It has been only relatively recently that word counts have become easy to display visually in word or tag clouds through popular online applications such as Wordle (http://wordle.net) or TagCrowd (http://tagcrowd.com), and word cloud tools have been added to many QDAS packages including NVivo, ATLAS.ti, Dedoose, and MAXQDA (see Appendix D). Although word cloud software creates dramatic visuals, significant concerns have been brought up about its use. One concern is that word clouds rely entirely on word frequency and do not provide context for readers to understand how words are used within a text (Harris, 2011). Word clouds are unable to differentiate between words with positive or negative connotations, and they can be visually

misleading because longer words take more space within the cloud (Viégas & Wattenberg, 2008, p. 51). Despite these concerns, word clouds' ease of use can make them a practical tool for social scientists if they are used sparingly and their limitations are acknowledged. Although not very useful for complex analysis, they can be used in a project's early phases to help researchers identify keywords in texts or compare multiple corpora or documents (Weisgerber & Butler, 2009). For example, two or more word clouds can be shown together to contrast word usage across corpora or documents (e.g., Uprichard, 2012). Advanced word cloud visualizations such as parallel tag clouds (Collins, Viégas, & Wattenberg, 2009) and SparkClouds (Lee, Riche, Karlson, & Carpendale, 2010) are recent developments that allow users to compare multiple word clouds. Finally, when coupled with written analysis and explanation, word clouds can be used to illustrate ideas or themes for lay audiences.

Word Trees and Phrase Nets

The two main tools for visualizing texts in terms of sentences and short phrases (rather than single words as in word clouds) are word trees and phrase nets. These tools were originally developed as part of IBM's project Many Eyes but are now available in NVivo and other QDAS packages. Word tree software allows researchers to see how a particular word is used in sentences or phrases and provides visual displays of the connection of an identified word or words to other words in a corpus through a branching system (Viégas & Wattenberg, 2008). These systems allow the researcher to have the tree branch to words that come either before or after the identified word, providing some context for words, which is an improvement over word clouds. For example, Henderson and Segal (2013) examined the relationship between a research university and local community organizations and found that the understandings and goals of research varied for the two groups. A word tree created from the study's documents displayed all the sentences that contain the word *research* to provide a better understanding of how this word was used and the variation of its use. Although they resemble word trees, phrase nets differ from word trees in that they focus on connections of word pairs rather than whole sentences (see van Ham, Wattenberg, & Viégas, 2009).

Although sentence visualization tools provide more contextual information than single word analysis, they are best suited for exploratory data analysis (Weisgerber & Butler, 2009) rather than for complex analysis or hypothesis testing. By focusing on keywords within sentences, word trees and phrase maps allow social scientists to quickly identify patterns of word use within corpora and whether words are being used in

divergent ways within or across texts.

Matrices and Maps

Matrices and maps are tools for the visualization of themes (rather than words or sentences) in texts. Since the process of identifying themes requires at least an initial analysis of the texts (see Chapter 11), visualization of themes is more valuable in the analysis and reporting phases of a project than in exploratory phases. Because researchers can rank themes or place them into nonordinal categories, visualizing a corpus at the thematic level offers more options and dimensions for visual representations than at the word or sentence level.

Matrices are sets of numbers arranged in rows or columns. In text analysis, a matrix involves “the crossing of two or more dimensions . . . to see how they interact” (Miles & Huberman 1994, p. 239). Matrices are very useful for organizing textual data and for visualizing the relationships between and among categories of data, examining how categories relate to theoretical concepts, and searching for propositions linking categories of data. Similar to a cluster heat map (Wilkinson & Friendly, 2009) or an ethnoarray (Dohan, Abramson, & Miller, 2012), a benefit of matrices is that they provide an overview of thematic patterns within corpora and allow for comparison between corpora. As with QDAS packages’ increased capabilities with word clouds, most qualitative software packages have the ability to create a matrix based on theme frequency with links to the corresponding text within the program. However, because matrices do not provide the stories or context behind the themes they organize, ideally, when creating matrices or other visualizations, researchers should link the individual boxes to quotes that support reader understanding of the theme.

Matrices, mind maps, and concept maps (Trochim, 1989; Wheeldon & Ahlberg, 2012) focus on the connections and relations between themes. In mind maps and concept maps, arrows indicate the direction of influence and can be made different thicknesses to signify the degree of connection if that information is available. Maps can be easily created for data analysis and reporting with standard or specialized software such as ATLAS.ti, MAXQDA, and NVivo. For example, Trochim, Cook, and Setze (1994) used concept mapping to develop a conceptual framework of the views of 14 staff members of a psychiatric rehabilitation agency’s views of a program of supported employment for individuals with severe mental illness. And Wheeldon and Faubert (2009) showed how concept maps can be used in data collection in an exploratory study of the perceptions of four Canadians.

Internet Resources

The Collaboration Site of Viégas and Wattenberg

http://hint.fm

“Visualizing the Future of Interaction Studies”

www.cios.org/ejcpublic/019/1/019125.HTML

The Word Tree, an Interactive Visual Concordance

http://hint.fm/papers/wordtree_final2.pdf

Wordle

www.wordle.net

TagCrowd

http://tagcrowd.com

Appendix H List of Websites

General Text Mining Websites

The DiRT Directory

http://dirtdirectory.org

The DiRT (Digital Research Tools) Directory is a registry of digital research tools for scholarly use. DiRT makes it easy for scholars conducting digital research to find and compare research tools including content management systems, optical character recognition software, statistical analysis packages, and visualization software.

Loughborough University's CAQDAS Site

www.restore.ac.uk/lboro/research/software/caqdas_comparison.php

This site offers a comparative overview of computer-assisted qualitative data analysis software (CAQDAS) packages (see Appendix D). The site is ordered by product functions rather than by software products.

The National Centre for Text Mining

www.nactem.ac.uk

The National Centre for Text Mining (NaCTeM) is the first publicly funded text mining center in the world. Operated by the University of Manchester, the site links to text mining services provided by NaCTeM, software tools, text mining groups seminars, general events, conferences, workshops, tutorials, demonstrations, and text mining publications.

The QDAS Networking Project

https://www.surrey.ac.uk/sociology/research/researchcentres/caqdas

This site, operated by the University of Surrey, provides practical support, training, and information in the use of a range of software programs designed to assist qualitative data analysis. It features platforms for debates concerning the methodological and epistemological issues arising from the use of such software packages.

Text Analysis Portal for Research

http://tapor.ca

The Text Analysis Portal for Research (TAPoR) is a gateway to the tools used in text mining and text analysis. The project is led by Rockwell, Sinclair, Uszkalo, and Radzikowska and housed at the University of Alberta. The site features software reviews and recommendations and links to papers, articles, and other sources of information about specific software tools.

Social Science Ethics Websites

Ethical Decision-Making and Internet Research: Recommendations From the AoIR Ethics Working Committee

http://aoir.org/reports/ethics2.pdf

The American Psychological Association Report Psychological Research Online: Opportunities and Challenges

www.apa.org/science/leadership/bsa/internet/internet-report.aspx

The British Psychological Society's Ethics Guidelines for Internet-Mediated Research

www.bps.org.uk/system/files/Public%20files/inf206-guidelines-for-internet-mediated-research.pdf

The Davis–Madsen Ethics Scenarios From the Academy of Management Blog Post "Ethics in Research Scenarios: What Would YOU Do?"

http://ethicist.aom.org/2013/02/ethics-in-research-scenarios-what-would-you-do

The Ethicist Blog From the Academy of Management

http://ethicist.aom.org

The Office of Research Integrity, U.S. Department of Health and Human Services

http://ori.hhs.gov

Social Science Writing Websites

The Social Science Writing Project

http://www.csun.edu/sswp

"What Is a Social Science Essay?"

www.sagepub.com/sites/default/files/upm-binaries/39896_9780857023711.pdf

"Becoming a 'Stylish' Writer: Attractive Prose Will Not Make You Appear Any Less Smart"

Rachel Toor
http://chronicle.com/article/Becoming-a-Stylish-Writer/132677

Open Access Journal Articles

“Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences”

Gregor Wiedemann
www.qualitative-research.net/index.php/fqs/article/view/1949
http://nbn-resolving.de/urn:nbn:de:0114-fqs1302231

“Hypertextuality, Complexity, Creativity: Using Linguistic Software Tools to Uncover New Information about the Food and Drink of Historic Mayans”

Rose Lema
www.qualitative-research.net/index.php/fqs/article/view/1852

“Text Mining Tools in the Humanities: An Analysis Framework”

Geoffrey Rockwell, John Simpson, Stéfan Sinclair, Kirsten Uszkalo, Susan Brown, Amy Dyrbye, and Ryan Chartier
http://journalofdigitalhumanities.org/2-3/text-mining-tools-in-the-humanities-an-analysis-framework

“Mapping Texts: Visualizing American Newspapers”

Andrew J. Torget and Jon Christensen
http://journalofdigitalhumanities.org/1-3/mapping-texts-project-by-andrew-torget-and-jon-christensen
http://mappingtexts.org

Appendix I Statistical Tools

Statistical analysis is traditionally performed on data organized in table (spreadsheet) format. In spreadsheets, data are arrayed in tables where each row represents a record and each column a variable, which is a feature of the record. Text mining data in tables contain two types of variables: (1) quantitative or numeric variables such as word frequency and (2) nominal or categorical variables such as codes for different words or phrases.

Statistical software packages such as STATA, SPSS, SAS, and R are often used to analyze data in tabular format from text mining projects. There are several statistical measurements and operations that are typically taught in social science statistics courses that are used in text mining and text analysis research, including reliability coefficients, analysis of variance (ANOVA), chi-square tests, and multiple regression (for general overviews, see Field, 2013; Field & Miles, 2012).

Reliability Coefficients

In statistics, interrater reliability (or interrater agreement) is the degree of agreement among raters. Interrater reliability is useful for determining if a particular scale is appropriate for measuring a particular variable; if multiple raters do not agree, either the scale is defective or the raters need retraining.

A number of statistics can be used to determine interrater reliability, with different statistics being appropriate for different types of measurement. Some options are joint probability of agreement, Cohen's kappa, Fleiss's kappa, interrater correlation, the concordance correlation coefficient, and intraclass correlation. But for text analysis and content analysis methods such as metaphor analysis, narrative analysis, and thematic analysis, the communications researcher and statistician Krippendorff's (2013, pp. 221–250) alpha coefficient is one of the most widely used statistical measure of interrater agreement. An advantage of Krippendorff's alpha is that it can deal with missing entries because it does not require the same number of raters for each item. Alpha can be calculated with software including SPSS and SAS (Hayes & Krippendorff, 2007) or in the statistics package R with the *kripp.alpha()* function in the interrater reliability package (https://www.rdocumentation.org/packages/irr/versions/0.70/topics/kripp.alpha).

Let's take a look at the use of reliability coefficients in the study of coverage of women's and men's sports in the *NCAA* (National Collegiate Athletic Association) *News*. In this 2004 study, the sport science researchers

Cunningham, Sagas, Satore, Amsden, and Schellhase (2004) had two coders independently code each of 5,745 paragraphs in their sample of magazine issues for gender and for the paragraph's location within the magazine and content. Prior to this coding process, three pilot tests had been conducted using previous issues not included in the main analyses. The pilot tests allowed the research team to clarify the definitions for each category to improve consistency in the rating process. The research team read randomly selected coded paragraphs out of the context of the article to ensure that the coding was consistent with the information in each paragraph. The researchers then used Cohen's kappa to estimate the reliability, which they found was high for the coding of the paragraphs' gender ($\kappa = .912$, $p < .001$), content ($\kappa = .964$, $p < .001$), and location ($\kappa = .997$, $p < .001$). The reliability as measured by the Pearson product moment correlation for length ($r = .995$) was also high. When differences in the coding did occur, the researchers met to discuss the differences until agreement was reached.

Analysis of Variance

Developed by the statistician and evolutionary biologist Fisher, ANOVA is a collection of statistical models used to analyze variation between groups. In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal. ANOVAs are useful for comparing the means of three or more groups or variables.

There are two main types of analysis of variance models: (1) fixed-effects and (2) random-effects models. The fixed-effects ANOVA model applies to situations in which an experimenter applies one or more treatments to the subjects of the experiment to see whether the response variable values change. Random-effects models are used when a study's various factor levels are sampled from a larger population.

Let's return to the study by Cunningham and his colleagues (2004) to understand why and how they used an ANOVA for comparing groups, as the goal of their study was to examine whether their text collection (issues of the magazine *NCAA News*) contained equitable distributions of content about men and women's sports teams. For such a study, a random-effects ANOVA makes sense, and Cunningham and colleagues (2004) used the ANOVA procedure to compare average paragraph length for stories about men's sports teams versus stories about women's teams. They did not find a significant difference between the two ($M = 2.25$, $SD = 2.17$ for men's teams; $M = 2.25$, $SD = 2.42$ for women's teams), $F(1, 4063) = .01$, $p = .94$, where M and SD are the mean and the standard deviation for each group, respectively, and the F statistic represents the

difference between the variance between groups (in this case men's versus women's sports teams) over the variance within groups in an ANOVA. Unlike in a t-test, the ANOVA procedure can be used to calculate the differences between groups larger than two. In this case, the p statistic of 0.94 shows that the F statistic is not statistically significant for the compared groups. In addition to comparing paragraph length, Cunningham and his colleagues (2004) also ran ANOVAs for gender differences in the size of the photographs featured in the magazine (see also Hirschman, 1987).

Chi-Square Tests

Where the ANOVA procedure is used to compare means across two or more groups, the chi-square statistic is used to compare word frequencies across documents or groups of documents that may differ in size. A statistical goodness-of-fit test originally suggested by the mathematician and biostatistician Pearson, the chi-square test is calculated based on the observed (actual) frequency Oi, the expected (averaged) frequency Ei, and the total frequency Ni in corpus i. The null hypothesis of the chi-square goodness-of-fit test is that there is no difference between the observed frequencies of a word in the two corpora. Even if the null hypothesis is not rejected, it cannot be concluded that it is true. The chi-square statistic is typically calculated on a 2 × 2 table to compare frequencies of words or other variables between two corpora.

Let's return to Cunningham and colleagues' (2004) study of the *NCAA News* for an example of the use of a chi-square test. Why did Cunningham and colleagues (2004) perform a chi-square test in addition to an ANOVA? First, they used a chi-square analysis to establish that the amount of coverage allotted to women's teams did not differ from 1999 to 2001, $\chi^2(1) = 3.65$, $p = .06$, where the p value of over 0.05 indicates that the difference in the *proportions* of coverage of women's to men's sports had not changed to a statistically significant degree. Where an ANOVA would allow for a comparison of means either across genders or over time, the chi-square allows for a comparison of the proportion of women to men for the two periods. These proportions can be easily visualized as a 2 × 2 table with gender on one axis and time on the other. In addition to their comparison of men's sports versus women's sports over time, Cunningham and colleagues (2004) compared their paragraphs in terms of information related to athletics versus information not related to athletics, for stories on men's teams and women's teams. Results indicate that paragraphs focused on women and women's teams were just as likely to contain information related to athletics (70.4%) as were paragraphs focused on men and men's teams (69.3%), $\chi^2(1) = .57$, $p = .45$. They also used a chi-square to compare the

proportion of women featured in photographs by year (see Ignatow, 2003, p. 12, for another example of the use of chi-square for comparing word frequency proportions).

Regression

Multiple regression is widely used in the social sciences to isolate the effects of one or more factors (*independent variables*) on some outcome of interest (a *dependent variable* or variables). Multiple regression, which is based on vector calculus, is a basic procedure featured in all statistical software packages used in the social sciences (see Field, 2013). In text mining and text analysis applications, regression, like ANOVA and chi-square, is used after data has been coded and univariate statistics (word frequencies and averages across groups) have been calculated. Regression models can be fitted to text data when the research question involves some factor, such as the speaker's age, gender, or number of friends, having an independent positive or negative effect on an outcome such as a frequency count or sentiment score (see Chapter 14 and Appendix E).

An example of the use of multiple regression is the study reviewed in Chapter 12 by management researchers Gibson and Zellmer-Bruhn. Gibson and Zellmer-Bruhn (2001) analyzed the relationship between metaphor use and employee attitudes in four countries. Their theoretical framework explained variance in the concept of teamwork across national and organizational cultures. Deriving five different metaphors for teamwork from team members' language used during interviews in four different geographic locations of six multinational corporations, they used the qualitative data analysis software QSR NUD*IST and TACT (see Appendix D) to analyze the frequency of the use of metaphors. For hypothesis testing, Gibson and Zellmer-Bruhn (2001) used a multinomial logit regression model (p. 293) where the dependent variable was the choice of teamwork metaphor from among the five possible types and the independent variables in their model were three dummy variables (binary variables that are coded either 1 [true] or 0 [false]) for country and five dummy variables for organization, representing the organizations in the study. Their models included control variables (independent variables that are not related to the research question or questions) for gender, team function, and the total number of words in each interview. Multiple regression allowed the researchers to control for (hold constant) gender, functional background, and total words in an interview. This revealed significant interaction effects that indicated that use of teamwork metaphors varied across countries and organizations net of other factors (for more detail, see Gibson & Zellmer-Bruhn, 2001, pp. 293–296).

Glossary

Abduction

A type of inferential logic in which the conclusion is a hypothesis that can then be tested with a new or modified research design, abduction is a forensic logic that is commonly use in social science research but also in natural science fields such as geology and astronomy where experiments are rarely performed.

Alceste

Software originally developed by Reinert in the 1980s, Alceste measures what Reinert termed *lexical worlds*, which he conceptualized as “mental rooms” that speakers successively inhabit, each with its own characteristic vocabulary.

Analogy

A form of metaphorical language that involves comparison between two things, typically on the basis of their correspondence or partial similarity

Analysis of variance (ANOVA)

Developed by the statistician and evolutionary biologist Fisher, analysis of variance (ANOVA) refers to statistical models for analyzing differences among group means and other statistics related to variation among groups.

Anonymize

In almost all text mining research, social scientists are required to anonymize (use pseudonyms for) users’ user names and full names.

Appendix

Located at the end of social science research papers, the appendices are optional sections, typically lettered A, B, C, etc., that contain information that may be useful to the reader but is not a critical component of the paper’s scientific contribution. An appendix may contain raw data, supplementary analysis, or other material.

Background metaphors

Widely used metaphors in a community or group that can be collected from sources such as encyclopedias,

journals, and specialist and generalist book according to Schmitt's (2005) method of comparative metaphor analysis

Bag of words

In topic modeling, the treatment of a text as combinations of word co-occurrences regardless of syntax, narrative, or location within a text

Bootstrapping

A list of organization names (or another named entity) and a list of patterns or rules to identify such organization names are incrementally learned from text

Case selection

Strategies and procedures used in ethnographic and historical research for selecting data sources

Characters

In narrative theory, characters are brought together with actions in a plotline that involves change over time.

Cognitive linguistics

The research field in which cognitive metaphor theory, which provides a conceptual foundation for most contemporary metaphor analysis methodologies, developed.

Cognitive metaphor theory (CMT)

Pioneered by cognitive linguists Lakoff and Johnson (1980), the basic claim of cognitive metaphor theory (CMT) is that language is structured by metaphor at a neural level and that metaphors used in natural language reveal cognitive schemas and associated patterns of neural connections shared by members of social groups.

Coherence theory

A major philosophical position that has influenced social science in which truth, knowledge, and theory must fit within a coherent system of propositions that may be applied to specific cases only in accordance with properties of the general system

Collocation

Closely related to linguistic co-occurrence, in corpus linguistics collocation refers to a sequence of words or terms that co-occur more often than would be expected by chance.

Collocation identification

Automatically identifying sequences of words that have a special meaning when taken as a phrase

Conceptual metaphors

In cognitive metaphor theory, natural language is characterized by the presence of conventional metaphorical expressions organized around prototypical metaphors, which Lakoff and Johnson (1980) have referred to as conceptual metaphors.

Conclusion

The final section of a research paper (although before the references and appendices), the conclusion section summarizes the main points made in the paper and makes suggestions for future research.

Concordance(s)

A list of the principal words used in a text, listing every instance of each word with its immediate context

Constructionism

Philosophical position based on questioning belief in an external reality that emphasizes how different groups construct their beliefs (see Gergen, 2015)

Content analysis

Research method for using systematic and generally quantitative tools to make inferences from recorded human communication (see Krippendorff, 2013)

Contextual level

Text analysis conducted at this level of analysis involves focusing on the immediate social context in which texts are produced and received, including situational contexts and the characteristics of the texts' authors.

Conversation analysis

An approach to the study of social interaction that began with a focus on casual daily conversation but expanded to include task- and institution-centered interactions such as those occurring in offices, courts, and

educational settings

Co-occurrences

Often interpreted as an indicator of semantic proximity and related to, but distinct from, linguistic collocation, co-occurrence refers to the above-chance occurrence of two terms from a text corpus located in close proximity to each other in a certain order.

Correspondence theory

Traditional model of knowledge and truth associated with scientific positivism, correspondence theory considers that there exists a correspondence between truth and reality and that notions of truth and reality correspond with things that actually exist in the world.

Crawlers

Automatic processes that browse the web to collect data

Critical case

A case selected because of its strategic importance to the research question

Critical discourse analysis (CDA)

Based on Fairclough's (1995) concept of "intertextuality," which is the idea that people appropriate from discourses circulating in their social space whenever they speak or write, CDA is a qualitative text analysis method that involves seeking the presence of features from other discourses in the text or discourse to be analyzed.

Critical metaphor analysis

A qualitative methodology developed by Charteris-Black (2009, 2012, 2013) that draws on methodologies and perspectives developed in cognitive linguistics, corpus linguistics, and critical linguistics that has been used to examine metaphors in political rhetoric, press reporting, and religion

Critical realism

Combining the realism of correspondence theory with the sociocultural reflexivity of social constructionism, in critical realism, some objects are understood to be more socially constructed than others.

Data mining

Analyzing large quantities of data inductively in search of trends and patterns

Data sampling

Refers to statistical techniques for selecting and analyzing a representative subset of data from a larger data set to identify trends and patterns in the larger data set

Decision trees

Structures that look like flowcharts; each internal node represents a test on one of the features and the nodes represent classification decisions

Deduction

The form of inferential logic most closely associated with the scientific method, deduction starts with theoretical abstractions, derives hypotheses from those theories, and then sets up research projects that test the hypotheses on empirical data.

Deep learning

One of the newest branches of machine learning, it consists of algorithms that aim to learn high-level representations of the data that can be used for effective learning.

Dictionary

An alphabetical list of the words in a language, which may include information such as definitions, usage examples, etymologies, translations, etc.

Digital archives

Collections of digital information, such as newspaper archives or archives of digitized historical documents, that are often accessible online and generally compatible with text mining research methods

Disambiguation

A text mining process that involves the use of contextual clues to decide where words refer to one or another of their multiple meanings

Discourse analysis

Involves seeking the presence of features from other discourses in the text or discourse to be analyzed based on the idea that people appropriate from discourses circulating in their social space whenever they speak or write

Discourse positions

Typical discursive roles that people adopt in their everyday communication practices, the analysis of discourse positions is a (generally qualitative) way of linking texts to the social spaces in which they have emerged.

Discussion

The discussion section of a research paper is located after the results (findings) but before the conclusions. The discussion section typically includes consideration of the meaning and implications of the results relative to the hypotheses/predictions and main argument.

Emplotment

In narrative theory, the process of bringing together characters and actions into a plot that involves change over time

Entity extraction

A subtask of information extraction (IE) that targets the identification of instances of a specific type, including named types, such as people or locations, or general semantic types, such as animals or colors.

Enumeration

A first element in any sampling strategy, enumeration involves assigning numbers to or comprehensively listing the units within a population.

Epistemology

Involves assumptions made in social science research about the nature of knowledge

Ethical guidelines

Generally published by academic and professional associations as well as by universities, these are guidelines for ethical social science research that cover issues such as informed consent and privacy.

Extreme case

Cases chosen for research that are thought to reveal more information because they activate more actors and more basic mechanisms in the situation studied

Feature ablation

A way to compare the performance of different features, by running the classifier using one feature at a time (forward ablation) or removing from the full feature set one feature at a time (backward ablation)

Feature vector

Collections of such properties, used to represent an instance of an event

Feature weighting

A technique used to indicate the role played by individual features in a classifier

Features (or attributes)

Measurable properties of an event being observed

Foucauldian analysis

A text analysis methodology that involves referring to the discourses that circulate in the social space in which the text is produced and received

Free clause

In Labov's (1972) narrative theory, a clause within a narrative that does not have a temporal component and can therefore be moved freely within the text without altering the text's meaning (see also *Minimal narrative*)

Functional approach

An approach to narrative pioneered by the psychologist Bruner (1990), who argued that humans' ordering of experience occurs in two modes: (1) a *paradigmatic,* or *logico–scientific mode*, which attempts to fulfill the ideal of a formal, mathematical system of description and explanation, and (2) a *narrative mode* in which events' particularity and specificity as well as people's involvement, accountability, and responsibility in bringing about specific events are centrally important

General Inquirer project

A long-running, large-scale content analysis project housed at Harvard University that involves developing a lexicon attaching syntactic, semantic, and pragmatic information to part-of-speech tagged words

Grand theory

Highly abstract and formal social theories that are broad in scope

Grounded theory

Systematic theory developed inductively based on observations that are grouped into conceptual categories (see Bryant & Charmaz, 2010)

Heaps' law

Models the number of words in the vocabulary as a function of the corpus size

Hypotheses

In deductive research, a proposition or set of propositions set forth as an explanation for the occurrence of a group of phenomena

Idiographic approaches

Approaches to causal explanation that emphasize concrete sequences of events, thoughts, and actions that lead to specific outcomes

Indigenous categories

Local terms that are used in unfamiliar ways and that can provide insights into themes and subthemes of the community being investigated

Induction

Involves making inferences that take empirical data as their starting point and work upward to theoretical generalizations

Inference

The process of deriving conclusions reached on the basis of evidence and reasoning

Inference to the best explanation

A type of inferential reasoning closely related to abduction

Information extraction (IE)

The task of extracting structured information such as entities, events, or relations from unstructured data

Informed consent

A core principle of human research ethics, established in the aftermath of the Second World War, that requires that research subjects explicitly agree to be participants in a research project based on a comprehensive understanding of what will be required of them

Instance-based learning

Form of lazy learning that includes algorithms such as *k*-nearest neighbors (KNN) and kernel machines

Institutional review board (IRB)

A university committee that has been formally designated to review, approve, and monitor social science and biomedical research involving humans

Intellectual property (IP)

Referring to intellectual creations for which a monopoly is assigned to their designated owners by law, common types of intellectual property rights are trademarks, copyright, and patents.

Introduction

In social science research papers, the introduction section includes background information about the phenomenon under investigation, review of relevant research literature, and the paper's research question or questions.

Language models

Probabilistic representations of language

Latent Dirichlet allocation (LDA)

Based on the idea that every text within a text collection is akin to a bag of words produced according to a mixture of topics that the author or authors intended to discuss, latent Dirichlet allocation (LDA) is a statistical model of language introduced by Blei, Ng, and Jordan (2003) for topic modeling.

Latent semantic analysis (LSA)

Based on the similarity of meaning of words appearing in texts or passages, a probabilistic model used in topic modeling that presents words and texts using vector space modelling that compiles textual data into a term-by-document matrix, showing the weighted frequency of terms to represent the documents in the term space

Learning curve

A graphical representation of the increase in learning performance (*y*-axis) with the amount of training data (*x*-axis)

Lemmatization

The process of identifying the base form (or root form) of a word

Levels of analysis

A term used in the social sciences to point to the scope or scale of the social phenomenon or phenomena to be studied

Linguistic markets

Also known in sociolinguistics and sociology as linguistic marketplaces, this concept refers to the symbolic markets in which linguistic exchanges occur. It is generally assumed that more standard language has higher value (prestige) in linguistic markets than nonstandard language (based on accent, vocabulary, and other factors).

Literature review

Literature review is sometimes included within a paper's introduction but is often its own section. Rather than simply reviewing the history of previous research on a particular topic, a good literature review is structured so that it highlights how the study stands to contribute to the literature by resolving a contradiction, solving a puzzle, or opening a new line of inquiry.

Logico–scientific mode

In Bruner's (1990) functional theory of narrative, the mode of organizing experience that attempts to fulfill the ideal of a formal, mathematical system of description and explanation (see also *Paradigmatic mode*)

Machine learning

A field in artificial intelligence that has had a very significant impact on a large number of problems in a diverse set of domains, ranging from information management to linguistics to astrophysics and many others

MALLET (Machine Learning for Language Toolkit)

A popular Java-based package used for topic modeling in the social sciences and humanities

Meso theory

Less sweeping and abstract than grand theory, and more closely connected to the practice of empirical research, meso theory draws on empirically supported substantive theories and models.

Metaphors

While metaphorical language takes a number of grammatical forms, including analogy, simile, and synecdoche, in all cases it involves figures of speech that make implicit comparisons in which a word or phrase ordinarily used in one domain is applied in another.

Metatheory

Reflection on the role of theory within social science research

Methods

The methods section of a social science paper includes a discussion of the method of analysis chosen and performed by the researchers. It typically includes discussion of why the method chosen is the best choice among all available methods as well as the details of how the method was used to analyze the data used in the paper.

Minimal narrative

In Labov's (1972) narrative theory, any sequence of two clauses that are temporally ordered (see also *Free clause*)

Mixed methods

Research methodologies that include both quantitative and interpretive elements

Modality analysis

A mixed method of narrative analysis intended for cross-cultural and cross-linguistic comparative research that evaluates languages by analyzing modal clauses in multiple large collections of text in multiple languages in order to identify what activities the users of each language treat as possible, impossible, inevitable, or contingent (see Roberts, 2008)

Naive Bayes

A classification technique based on Bayes theorem

Named entity recognition (NER)

Refers to tools for identifying proper names that can be classified under a certain type

Narrative analysis

A form of qualitative analysis that focuses on how people tell stories to make sense of everyday experiences and events in their lives (see Holstein & Gubrium, 2011)

Narrative clauses

In the narratologist Labov's (1972) terminology, narrative clauses are clauses within a story that have a temporal component (rather than providing background information). Movements of narrative clauses with a story change the story's meaning.

Narrative mode

In Bruner's (1990) functional theory of narrative, the mode of organizing experience in which events' particularity and specificity as well as people's involvement, accountability, and responsibility in bringing about specific events are centrally important.

Natural language processing (NLP)

The process or ability of a machine or program to understand natural (or human) text or speech

Netnography

The use of ethnographic methods to study online communities (see also *Virtual ethnography*)

Network techniques

Text analysis methods that model statistical associations between words to infer the existence of mental

models shared by members of a community

_N_-fold cross-validation

A technique to evaluate a machine-learning classifier by partitioning the data into N folds and repeatedly training the classifier on (*N*−1) folds and testing it on the remaining fold

Nomothetic approaches

Approaches to causal explanation that emphasize common influences on a number of cases or events

Normalization

The process of transforming text into a canonical form (e.g., by expanding abbreviations, by correcting misspellings)

Ontology

Involves assumptions about the nature of reality

Opinion mining

The task of identifying such private states in language; it has two main subtasks: (1) subjectivity analysis and (2) sentiment analysis

OpinionFinder

Includes words and phrases that are indicators of subjectivity, along with a polarity markup

Paradigmatic mode

In Bruner's (1990) functional theory of narrative, the mode of organizing experience that attempts to fulfill the ideal of a formal, mathematical system of description and explanation (also referred to as the *logico–scientific mode*)

Part-of-speech tagging

The task of assigning each word in an input text with its correct syntactic role, such as noun, verb, and so forth

Password-protected data

Because users posting in password-protected websites are likely to have expectations of privacy, there is widespread agreement that websites that require registration and password-protected data should be considered to be in the private domain.

Philosophy of social science

An academic research area that lies at the intersection of philosophy and social science, philosophy of social science involves the development and critique of concepts that are foundational to the practice of social science research.

Plagiarism

A major ethical concern in social science research that involves the wrongful appropriation of another researcher's language or ideas and the representation of them as one's own original work

Plotline

In theories of narrative, the narrative's structure in which characters and actions are brought together and change over time

Pointwise mutual information (PMI)

A measure of (word) association stemming from information theory

Polarity label

Indicates whether the corresponding word or phrase is positive, negative, or neutral

Pragmatism

Approaches to the philosophy of social science in which truth is defined as those tenets that prove useful to the believer or user, and truth and knowledge are verified through experience and practice

Preprocessing

A sequence of basic text processing steps applied in advance of more complex processing, typically consisting of tokenization, lemmatization, and normalization

Privacy

A major ethical concern for text mining researchers that is treated somewhat differently in different national

contexts (e.g., in the European Union versus the United States) and in academic and professional associations' ethical guidelines

Probability sample

Involves taking someone else's research or ideas and passing them off as one's own

Prompted data

Collecting users' textual data after actively manipulating the online environment as a stimulus intended to assess reactions or responses

Public domain

Data in the public domain can be used freely by text mining researchers, although many websites and social media platforms have privacy policies that set expectations for users' privacy and can be used as guidelines for whether it is ethical to treat the site's data as in the public domain or whether informed consent may be required.

Public stories

Narratives that circulate in popular culture

Purposive sampling (see also Relevance sampling)

A research-question-driven, nonprobabilistic sampling technique in which the researcher learns about a population of interest and then gradually reduces the number of texts to be analyzed based on the texts' relevance to the research question

Qualitative analysis

Text analysis methodologies based on human interpretation of texts

Quantitative analysis

Text analysis methodologies based on mathematical and statistical techniques

Random sample

A sampling strategy that reduces sample bias by using a randomization device such as software or an online random number generator to select items of data from an enumerated data set

References

Reference sections of social science papers include all the publications (papers, books, book chapters, conference proceedings, websites) cited in the paper. Several different reference formats are widely used in the social sciences, including APA (American Psychological Association) and Chicago formats.

Registration

Websites that require user registration and password-protected data are generally considered to be in the private domain.

Regression (multiple regression)

A collection of statistical techniques used to predict the value of one variable (a “dependent” variable) based on the value of two or more other variables (“independent” variables)

Relation extraction

A subtask of information extraction (IE) that aims to identify relationships between entities, such as “capital of,” sibling, and so forth

Relevance sampling (see Purposive sampling)

A research question-driven, nonprobabilistic sampling technique in which the researcher learns about a population of interest and then gradually reduces the number of texts to be analyzed based on the texts’ relevance to the research question

Repeated reading

This is the first step in thematic analysis where the researcher acquires a collection of texts and reads them repeatedly while searching for themes and taking extensive notes (see Braun & Clarke, 2006).

Representative case

Data that are representative of a larger population that are selected when the objective of a project is to achieve the greatest possible amount of information about a phenomenon

Research design

The phase of research concerned with the basic architecture of research projects and designing projects that

allow theory, data, and method to interface in such a way as to maximize a project's ability to achieve its goals

Results

The results section includes the results of the analysis performed. Results are presented in a straightforward manner, generally with a high level of technical detail but little interpretation of their meaning.

Rocchio classifier

Builds upon the ideas of the vector-space model used in information retrieval

Sample bias

A major disadvantage of Internet-accessed data samples is that it is very difficult to draw inferences about the larger population that the sampled data are claimed to represent. This is due to sample bias based on factors such as level of Internet access, level of Internet skill, and specific characteristics (such as comment moderation strategies) of websites and social media platforms.

Sampling

Involves the selection of a subset of data from within a statistical population to estimate characteristics of the whole population

Scientific method

A formal method of research that involves identification of a problem, collection of relevant data, formulation of hypotheses, and empirical testing of hypotheses

Scrapers

Automatic processes for extracting data from websites

Semantic networks

A network that defines the semantic relations between words

Semantic relations

Relations that exist between word meanings

Semantic techniques

Sometimes referred to as hermeneutic or hermeneutic structuralist techniques, these include a variety of methods designed to recognize latent meanings in texts.

Semantic triplet

In Franzosi's approach to narrative, a fundamental semantic structure involving an actor, action, and object of action

Sentiment analysis

Use of software to discern subjective material and extract various forms of attitudinal information such as sentiment, opinion, mood, and emotion

SentiWordNet

A resource for opinion mining built on top of WordNet, which assigns each synset in WordNet with a score triplet, indicating the strength of each of these three properties for the words in the synset

Simile

A form of metaphorical language that involves the comparison of one thing with another thing of a different kind, used to make a description more emphatic or vivid

Snowball sampling

A widely used iterative sampling technique in which a researcher starts with a small sample and then repeatedly applies a sampling criterion until a sample size limit is reached

Sociological approaches

Approaches to narrative that focus on the cultural, historical, and political contexts in which particular stories are, or can be, told by particular narrators to particular audiences

Sociological level

A level of analysis at which attempts are made to identify causal relations between texts and the social contexts in which they are produced and received

Source domain

In cognitive metaphor theory, perceptual and sensory experiences from an embodied source domain, such

as pushing, pulling, supporting, balance, straight–curved, near–far, front–back, and high–low, are used to represent abstract entities in a target domain.

Stemming

A processing step that uses a set of rules to remove inflections

Story grammar

In structural narrative analysis, a basic narrative structure that is repeated across many diverse narrative genres

Stratified sampling

A sampling strategy that involves sampling from within subunits (“strata”) of a population

Structural approaches

Approaches to narrative analysis pioneered by Propp (1968) and Labov (1972) that center on story grammars and other basic structural features of narratives that are found in narratives from diverse sources

Subjectivity analysis

Identifies if a text contains an opinion and correspondingly labels the text as either subjective or objective

Substantive theory

Theory derived from data analysis that involves rich conceptualizations of specific social and historical situations

Supervised learning

Consists of using an automatic system to learn from a history of occurrences of a certain “event” and consequently make predictions about future occurrences of that event

Support vector machines (SVMs)

Supervised learning machine algorithms that identify the hyperplane that best separates the training data

Synecdoche

A form of metaphorical language involving a figure of speech in which a part is made to represent the whole

or vice versa

Syntactic parsing

Within computational linguistics, syntactic parsing refers to the use of software to formally analyze a sentence or other string of words in terms of its constituents, resulting in a parse tree showing words' and phrases' syntactic relations to each other.

Systematic sampling

A sampling strategy that involves sampling every *k*th unit from an enumerated list

Target domain

In cognitive metaphor theory, the target domain is the relatively abstract or complex entity that is represented by perceptual and sensory experiences drawn from an embodied source domain. For instance, the relatively abstract concept of *argument* can be a target domain where *battle* would be a possible source domain. In this way, *argument* is understood to have many of the qualities (e.g., the existence of a winner and a loser) of a battle.

Template filling

The overall process of filling in the values for all the aspects in a template

Text analysis

In the social sciences, text analysis refers to methods of systematically analyzing word use patterns in texts that often combine formal statistical methods and less formal, more humanistic interpretive techniques.

Text classification

The process of assigning texts to one or more predefined categories

Text clustering

The process of grouping texts into clusters of texts based on their similarity

Text mining

The use of digital research tools to derive high-quality information from textual data

Textual level

A level of analysis at which attempts are made to characterize or determine the composition and structure of the text itself

Thematic analysis

A method of text analysis for identifying, analyzing, and reporting patterns of themes within texts (see Boyatzis, 1998)

Thematic coding

Within thematic analysis, the process of systematically labeling texts based on predetermined or emergent categories

Thematic techniques

Text analysis techniques that are focused on manifest meanings in texts and include methods commonly used in business as well as the social sciences such as topic modeling

Theoretical models

Simplified, often schematic representations of complex social phenomena that are used in almost all social science research but particularly in research that is done in a positivist mode of inquiry

Thesauruses

Databases that group the words in a language according to similarities

Tokenization

The process of separating the punctuation from the words while maintaining their intended meaning

Topic models

Involve automated procedures for coding collections of texts in terms of meaningful categories that represent the main topics being discussed

Transformation

In narratology, transformation refers to changes in characters over time that result from events and characters'

actions.

Traversal strategies

Methods that define the sequence of steps that a web crawler will take; typical traversal strategies are depth-first and breadth-first

Units of analysis

For texts, the unit of analysis can be conceived in many ways, including in terms of hierarchies in which one level includes the next, or as sequentially ordered events, or as networks of intertextual relationships.

Unstructured data

Free-form text data are considered to be unstructured data because they are not organized in a predefined manner (such as in a matrix with rows and columns).

Unsupervised learning

A type of machine-learning algorithm that makes inferences using unlabeled data

URL

The address of a webpage

Varying probability sampling

A sampling strategy used to sample proportionately from data sources with different sizes or levels of importance, such as newspapers with different circulation levels

“Verbed”

It refers to the widespread tendency in academic writing to turn nouns into unfamiliar and sophisticated-sounding verbs. It should be avoided whenever possible.

Virtual ethnography

The use of ethnographic methods to study online communities (see also *Netnography*)

Web crawling

Use of Internet bots to systematically browse the World Wide Web for the purpose of web indexing

Web (open) information extraction (IE)

A recently defined task that aims to perform information extraction (IE) at scale, without the need to predefine the entities or relationships that are being extracted

Web scraping

A computer software technique of extracting information from websites, usually with programs that simulate human exploration of the World Wide Web

Word sense disambiguation

Maps input words to dictionary senses and is used to identify the meaning of a word as a function of its context

Word similarity

A measure that reflects the semantic closeness between two words

WordNet-Affect

Another extension of WordNet, it includes affective labels assigned to word senses; the six emotion categories of specific interest are (1) anger, (2) disgust, (3) fear, (4) joy, (5) sadness, and (6) surprise.

Zipf's law

Models the distribution of terms in a corpus and provides a mathematical way to answer this question: How many times does the *r*th most frequent word appear in a corpus of N words?

Zombie nouns

Often formed with the suffix *-ism*, zombie nouns are nouns formed from other parts of speech such as adjectives and verbs. Like "verbed," the formation of zombie nouns should be avoided.

References

Acerbi, A., Lampos, V., Garnett, P., & Bentley, A. (2013, March20). *The expression of emotions in 20th century books*. *PLOS ONE*. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0059030

Adams, J. (2009). Bodies of change: A comparative analysis of media representations of body modification practices. *Sociological Perspectives*, 52(1), 103–129.

Adams, J., & Roscigno, V. (2005). White supremacists, oppositional culture and the World Wide Web. *Social Forces*, 84(2), 759–778.

Albergotti, R., & Dwoskin, E. (2014, June30). Facebook study sparks soul-searching and ethical questions. *Wall Street Journal*.

Alder, K. (2007). *The lie detectors: The history of an American obsession*. New York, NY: Simon & Schuster.

Alm, C. O., Roth, D., & Sproat, R. (2005). *Emotions from text: Machine learning for text-based emotion prediction*. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 579–586). Stroudsburg, PA: Association for Computational Linguistics.

Andersen, D. (2015). Stories of change in drug treatment: A narrative analysis of “whats” and “hows” in institutional storytelling. *Sociology of Health & Illness*, 37(5), 668–682.

Asher, K., & Ojeda, D. (2009). Producing nature and making the state: Ordenamiento territorial in the Pacific lowlands of Colombia. *Geoforum*, 40(3), 292–302.

Asplund, T. (2011). Metaphors in climate discourse: An analysis of Swedish farm magazines. *Journal of Science Communication*, 10(4), 1–10.

Attard, A., & Coulson, N. (2012). A thematic analysis of patient communication in Parkinson’s disease online support group discussion forums. *Computers in Human Behavior*, 28(2), 500–506.

Ayers, E. L. (1999). *The pasts and futures of digital history*. Retrieved June17, 2015, from http://www.vcdh.virginia.edu/PastsFutures.html

Bail, C. (2012). The fringe effect: Civil society organizations and the evolution of media discourse about Islam since the September 11th attacks. *American Sociological Review*, 77(6), 855–879.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., Mcenery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses

of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306.

Balog, K., Mishne, G., & de Rijke, M. (2006). *Why are they excited? Identifying and explaining spikes in blog mood levels*. *Proceedings of the Eleventh Meeting of the European Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.

Bamberg, M. (2004). Form and functions of “slut bashing” in male identity constructions in 15-year-olds. *Human Development*, 47(6), 331–353.

Banerjee, S., & Pedersen, T. (2002). *An adapted Lesk algorithm for word sense disambiguation using WordNet*. Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007, January). Open information extraction from the web. *Communications of the ACM—Surviving the Data Deluge*, 51(12), 68–74.

Bastin, G., & Bouchet-Valat, M. (2014). Media corpora, text mining, and the sociological imagination: A free software text mining approach to the framing of Julian Assange by three news agencies. *Bulletin de Méthodologie Sociologique*, 122, 5–25.

Bauer, M. W., Bicquelet, A., & Suerdem, A. K. (Eds.). (2014). *Text analysis: An introductory manifesto*. In M. W.Bauer, A.Bicquelet, & A. K.Suerdem (Eds.), *Textual analysis (SAGE benchmarks in social research methods)* (Vol. 1). Thousand Oaks, CA: Sage.

BauerM. W., Gaskell, G., & Allum, N. (2000). *Quantity, quality and knowledge interests: Avoiding confusions*. In M. W.Bauer & G.Gaskell (Eds.), *Qualitative researching with text, image and sound* (pp. 3–17). Thousand Oaks, CA: Sage.

Becker, H. S. (1993). How I learned what a crock was. *Journal of Contemporary Ethnography*, 22, 28–35.

Bednarek, M., & Caple, H. (2014). Why do news values matter? Towards a new methodological framework for analyzing news discourse in critical discourse analysis and beyond. *Discourse & Society*, 25(2), 135–158.

Beer, F. A., & De Landtsheer, C. L. (2004). *Metaphorical world politics: Rhetorics of democracy, war and globalization*. East Lansing: Michigan State University.

Bell, E., Campbell, S., & Goldberg, L. R. (2015). Nursing identity and patient-centredness in scholarly health services research: A computational text analysis of PubMed abstracts, 1986–2013. *BMC Health Services Research*, 15(3), 1–16.

Berelson, B. (1952). *Content analysis in communication research*. Glencoe, IL: Free Press.

Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. Garden City, NY: Doubleday.

Berglund, E. (2001). Facts, beliefs and biases: Perspectives on forest conservation in Finland. *Journal of Environmental Planning and Management*, 44, 833–849.

Bernard, R., Wutich, A., & Ryan, G. (2016). *Analyzing qualitative data: Systematic approaches*. Thousand Oaks, CA: Sage.

Berry, M., Dumais, S., & O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4), 573–595.

Bhaskar, R. (2008). *A realist theory of science*. New York, NY: Routledge. (Original work published 1975)

Bickes, H., Otten, T., & Weymann, L. C. (2014). The financial crisis in the German and English press: Metaphorical structures in the media coverage on Greece, Spain and Italy. *Discourse & Society*, 25(4), 424–445.

Bicquelet, A., & Weale, A. (2011). Coping with the cornucopia: Can text mining help handle the data deluge in public policy analysis?*Policy and Internet*, 3(4), 1–21.

Biernacki, R. (2014). Humanist interpretation versus coding text samples. *Qualitative Sociology*, 37(2), 173–188.

Birke, J., & Sarkar, A. (2007). *Active learning for the identification of nonliteral language*. Proceedings of the Workshop on Computational Approaches to Figurative Language, 21–28.

Birnbaum, M. H. (2000). *Decision making in the lab and on the web*. In M. H.Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 3–34). Cambridge, MA: Academic Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Blevins, C. (2011, June19–22). *Topic modeling historical sources: Analyzing the diary of Martha Ballard*. *Digital Humanities*, Stanford University, Stanford, CA. Retrieved from http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-173.xml;query=;brand=default

Bolden, R., & Moscarola, J. (2000). Bridging the quantitative-qualitative divide: The lexical approach to textual data analysis. *Social Science Computer Review*, 18(4), 450–460.

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75(1), 1–28.

Boussalisa, C., & Coan, T. G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change*, 36, 89–100.

Bourdieu, P., & Thompson, J. B. (1991). *Language and symbolic power*. Cambridge, MA: Harvard University Press.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA: Sage.

Bradley, J. (1989). *TACT user manual*. Toronto, Canada: University of Toronto Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.

Brill, E. (1992). *A simple rule-based part of speech tagger*. Proceedings of the Third Conference on Applied Natural Language Processing. Trento, Italy.

Broaddus, M. (2014, July1). *Issues of research ethics in the Facebook "Mood Manipulation" Study: The importance of multiple perspectives*. *Ethics and Society*. Retrieved from https://ethicsandsociety.org/2014/07/01/issues-of-research-ethics-in-the-facebook-mood-manipulation-study-the-importance-of-multiple-perspectives-full-text

Brugidou, M. (2003). Argumentation and values: An analysis of ordinary political competence via an open-ended question. *International Journal of Public Opinion Research*, 15(4), 413–430.

Brugidou, M., Escoffier, C., Folch, H., Lahlou, S., Le Roux, D., Morin-Andréani, P., & Piat, G. (2000). *Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles. The factors of choice and use of textual data analysis software*. In *JADT 2000* (5èmes Journées Internationales d'Analyse Statistique des Données Textuelles).

Bruner, J. S. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

Bryant, A., & Charmaz, K. (Eds.). (2010). *The SAGE handbook of grounded theory*. Thousand Oaks, CA: Sage.

Buchholz, M. B., & von Kleist, C. (1995). *Psychotherapeutische Interaktion—Qualitative Studien zu Konversation und Metapher, Geste und Plan*. Opladen: Westdeutscher Verlag.

Bunn, J. (2012). *The truth machine: A social history of the lie detector (Johns Hopkins studies in the history of technology)*. Baltimore, MD: Johns Hopkins University Press.

Busanich, R., McGannon, K., & Schinke, R. (2014). Comparing elite male and female distance runners' experiences of disordered eating through narrative analysis. *Psychology of Sport and Exercise*, 15(6), 705–712.

Cameron, L. (2003). *Metaphor in educational discourse*. New York, NY: Continuum.

Carenini, G., Ng, R., & Zhou, X. (2007). *Summarizing emails with conversational cohesion and subjectivity*. Proceedings of the Sixteenth International Conference on World Wide Web. New York, NY: Association for Computing Machinery.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., Jr., & Mitchell, T. M. (2010, July). *Toward an architecture for never-ending language learning*. Proceedings of the Twenty-Fourth American Association for Artificial Intelligence Conference on Artificial Intelligence (pp. 1306–1313). Cambridge, MA: AAAI Press.

Carver, T., & Pikalo, J. (2008). *Political language and metaphor: Interpreting and changing the world*. New York, NY: Routledge.

Cerulo, K. A. (1998). *Deciphering violence: The cognitive structure of right and wrong*. New York, NY: Routledge.

Chalaby, J. K. (1996). Beyond the prison-house of language: Discourse as a sociological concept. *The British Journal of Sociology*, 47(4), 684–698.

Chambers, C. (2014, July1). *Facebook fiasco: Was Cornell's study of "emotional contagion" an ethics breach?Guardian*. Retrieved from https://www.theguardian.com/science/head-quarters/2014/jul/01/facebook-cornell-study-emotional-contagion-ethics-breach

Charteris-Black, J. (2009). *Metaphor and political communication*. In A.Musolff & J.Zinken (Eds.), *Metaphor and discourse* (pp. 97–115). Basingstoke, England: Palgrave Macmillan.

Charteris-Black, J. (2012). *Comparative keyword analysis and leadership communication: Tony Blair—A study of rhetorical style*. In L.Helms (Ed.), *Comparative political leadership* (pp. 142–164). Basingstoke, England: Palgrave Macmillan.

Charteris-Black, J. (2013). *Analysing political speeches: Rhetoric, discourse and metaphor*. Basingstoke, England: Palgrave Macmillan.

Chilton, P. (1996). *Security metaphors: Cold War discourse from containment to common house*. New York, NY: Peter Lang.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.

Coffey, A., Holbrook, B., & Atkinson, P. (1996). *Qualitative data analysis: Technologies and representations*. *Sociological Research Online*, 1(1). Retrieved from http://www.socresonline.org.uk/1/1/4.html

Cohen, D. J., & Rosenzweig, R. (2005). *Digital history: A guide to gathering, preserving, and presenting the past on the web*. Philadelphia: University of Pennsylvania Press.

Collins, C., Viégas, F. B., & Wattenberg, M. (2009). *Parallel tag clouds to explore and analyze faceted text corpora*. *IEEE Symposium on Visual Analytics Science and Technology*. Retrieved June17, 2015, from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5333443&tag=1

Collins, M. (2002). *Ranking algorithms for named-entity extraction: Boosting and the voted perceptron*. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 489–496). Stroudsburg, PA: Association for Computational Linguistics.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4), 589–637.

Collins, M., & Singer, Y. (1999). *Unsupervised models for named entity classification*. Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Colley, S. K., & Neal, A. (2012). Automated text analysis to examine qualitative differences in safety schema among upper managers, supervisors and workers. *Safety Science*, 50(9), 1775–1785.

Corley, P., Collins, Jr., P., & Calvin, B. (2011). Lower court influence on U.S. Supreme Court opinion content. *Journal of Politics*, 73(1), 31–44.

Coulson, N. S. (2005). Receiving social support online: An analysis of a computer-mediated support group for individuals living with irritable bowel syndrome. *CyberPsychology & Behavior*, 8(6), 580–584.

Coulson, N. S., Buchanan, H., & Aubeeluck, A. (2007). Social support in cyberspace: A content analysis of communication within a Huntington's disease online support group. *Patient Education and Counseling*, 68(2),

173–178.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64(4), 464–494.

Creswell, J. D. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.

Cunningham, G. B., Sagas, M., Sartore, M. L., Amsden, M. L., & Schellhase, A. (2004). Gender representation in the *NCAA News*: Is the glass half full or half empty?*Sex Roles*, 50(11–12), 861–870.

Curd, M., Cover, J. A., & Pincock, C. (2013). *Philosophy of science: The central issues* (2nd ed.). New York, NY: W. W. Norton.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012, April16–20). *Echoes of power: Language effects and power differences in social interaction*. *WWW 2012*. Retrieved March29, 2016, from http://www.cs.cornell.edu/~cristian/Echoes_of_power_files/echoes_of_power.pdf

Danesi, M. (2012). *Linguistic anthropology: A brief introduction*. Toronto: Canadian Scholars' Press.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6), 391–407.

Denzin, N. K, & Lincoln, Y. S. (2011). *Epilogue: Toward a "refunctioned ethnography."The SAGE Handbook of Qualitative Research* (pp. 715–718). Thousand Oaks, CA: Sage.

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

Dohan, D., Abramson, C. M., & MillerS. (2012). *Beyond text: Using arrays of ethnographic data to identify causes and construct narratives*. Presentation at the American Journal of Sociology Conference on Causal Thinking and Ethnographic Research. Chicago, IL.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1), 188–230.

Durland, M., & Fredericks, K. (2005). An introduction to social network analysis. *New Directions for Evaluation*, 107, 5–13.

Edley, N., & Wetherell, M. (1997). Jockeying for position: The construction of masculine identities. *Discourse*

& Society, 8(2), 203–217.

Edley, N., & Wetherell, M. (2001). Jekyll and Hyde: Men's construction of feminism and feminists. *Feminism & Psychology*, 11(4), 439–457.

Ensslin, A., & Johnson, S. (2006). Language in the news: Investigating representations of "Englishness" using WordSmith tools. *Corpora*, 1(2), 153–185.

Eshbaugh-Soha, M. (2010). The tone of local presidential news coverage. *Political Communication*, 27(2), 121–140.

Esuli, A., & Sebastiani, F. (2006a). *Determining term subjectivity and term orientation for opinion mining*. Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy.

Esuli, A., & Sebastiani, F. (2006b). *SentiWordNet: A publicly available lexical resource for opinion mining*. Proceedings of the Fifth Conference on Language Resources and Evaluation, Genova, Italy.

Etzioni, O., Cafaraella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., . . . Yates, A. (2004). *Web-scale information extraction in KnowItAll: (Preliminary results)*. Proceedings of the Thirteenth International Conference on World Wide Web (pp. 100–110). New York, NY: Association for Computing Machinery.

Evison, J. (2013). Turn openings in academic talk: Where goals and roles intersect. *Classroom Discourse*, 4(1), 3–26.

Eysenbach, G., & Till, J. E. (2001). Ethical issues in qualitative research on Internet communities. *British Medical Journal*, 323, 1103–1105.

Fader, A., Soderland, S., & Etzioni, O. (2011). *Identifying relations for open information extraction*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1535–1545). Stroudsburg, PA: Association for Computational Linguistics.

Fairclough, N. (1992). Intertextuality in critical discourse analysis. *Science Direct*, 4(3–4), 269–293.

Fairclough, N. (1995). *Critical discourse analysis: The critical study of language*. London, England: Longman.

Fass, D. (1991). Met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1), 49–90.

Feldman, J. (2006). *From molecule to metaphor*. Cambridge, MA: MIT Press.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fenton, F. (1911). The influence of newspaper presentations upon the growth of crime and other anti-social activity. *American Journal of Sociology*, 16(3), 342–371.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1), 80–92.

Fernandez, J. W. (1991). *Beyond metaphor: The theory of tropes in anthropology*. Stanford, CA: Stanford University Press.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Thousand Oaks, CA: Sage.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Thousand Oaks, CA: Sage.

Flyvbjerg, B. (2001). *Making social science matter: Why social inquiry fails and how it can succeed again*. Cambridge, England: Cambridge University Press.

Fors, A., Dudas, K., & Ekman, I. (2014). Life is lived forwards and understood backwards—Experiences of being affected by acute coronary syndrome: A narrative analysis. *International Journal of Nursing Studies*, 51(3), 430–437.

Foucault, M. (1973). *The order of things: An archaeology of the human sciences*. New York, NY: Vintage Books.

Foucault, M. (1975). *The birth of the clinic: An archaeology of medical perception*. New York, NY: Vintage Books.

Franklin, S. (2002). Bialowieza Forest, Poland: Representation, myth, and the politics of dispossession. *Environment and Planning*, 34, 1459–1485.

Franzosi, R. (1987). The press as a source of socio-historical data: Issues in the methodology of data collection from newspapers. *Historical Methods*, 20(1), 5–16.

Franzosi, R. (2010). *Quantitative narrative analysis*. Thousand Oaks, CA: Sage.

Franzosi, R., De Fazio, G., & Vicari, S. (2012). Ways of measuring agency: An application of quantitative narrative analysis to lynchings in Georgia (1875–1930). *Sociological Methodology*, 42(1), 1–42.

Freud, S. (2011). *From the history of an infantile neurosis—A classic article on psychoanalysis*.

Worcestershire, England: Read Books. (Original work published 1918)

Frith, H., & Gleeson, K. (2004). Clothing and embodiment: Men managing body image and appearance. *Psychology of Men & Masculinity*, 5(1), 40–48.

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

Gandy, L., Allan, N., Atallah, M., Frieder, O., Howard, N., Kanareykin, S., . . . Argamon, S. (2013). *Automatic identification of conceptual metaphors with limited knowledge*. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. Bellevue, Washington.

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). *Studying online social networks*. *Journal of Computer Mediated Communication*, 3(1) http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1997.tb00062.x/abstract.

Gatti, L., & Catalano, T. (2015). The business of learning to teach: A critical metaphor analysis of one teacher's journey. *Teaching and Teacher Education*, 45, 149–160.

Gee, J. P. (1991). A linguistic approach to narrative. *Journal of Narrative and Life History*, 1(1), 15–39.

Gergen, K. (2015). *An invitation to social construction*. Thousand Oaks, CA: Sage.

Gerrish, S., & Blei, D. (2012). *How they vote: Issue-adjusted models of legislative behavior*. *Neural Information Processing Systems*. Retrieved June26, 2015, from https://www.cs.princeton.edu/~blei/papers/GerrishBlei2012.pdf

Gibbs, R. W. (1994). *The poetics of mind: Figurative thought, language, and understanding*. Cambridge, England: Cambridge University Press.

Gibson, C. B., & Zellmer-Bruhn, M. E. (2001). Metaphors and meaning: An intercultural analysis of the concept of teamwork. *Administrative Science Quarterly*, 46(2), 274–303.

Glaser, B., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Piscataway, NJ: Transaction Publishers.

Goatly, A. (2007). *Washing the brain: Metaphor and hidden ideology*. Philadelphia, PA: John Benjamins Publishing Company.

Goble, E., Austin, W., Larsen, D., Kreitzer, L., & Brintnell, E. S. (2012). *Habits of mind and the split-mind effect: When computer-assisted qualitative data analysis software is used in phenomenological research*.

Forum: Qualitative Social Research, 13(2). Retrieved June26, 2015, from http://www.qualitative-research.net/index.php/fqs/article/view/1709

Goldstone, A., & Underwood, T. (2012). *What can topic models of PMLA teach us about the history of literary scholarship?The Stone and the Shell*. Retrieved June27, 2015, from tedunderwood.com/2012/12/14/what-can-topic-models-of-pmla-teach-us-about-the-history-of-literary-scholarship

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). *Identifying sarcasm in Twitter: A closer look*. *Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies—Short Papers Volume 2*. Stroudsburg, PA: Association for Computational Linguistics.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Gorard, S. (2013). *Research design: Creating robust approaches for the social sciences*. Thousand Oaks, CA: Sage.

Gorbatai, A., & Nelson, L. (2015). *The narrative advantage: Gender and the language of crowdfunding*. Retrieved from http://faculty.haas.berkeley.edu/gorbatai/working%20papers%20and%20word/Crowdfunding-GenderGorbataiNelson.pdf

Gorski, D. (2014, June30). *Did Facebook and PNAS violate human research protections in an unethical experiment?Science-Based Medicine*. Retrieved from https://sciencebasedmedicine.org/did-facebook-and-pnas-violate-human-research-protections-in-an-unethical-experiment

Gottschall, J. (2012). *The storytelling animal*. New York, NY: Houghton Mifflin.

Gregory, M., Chinchor, N., Whitney, P., Carter, R., Hetzler, E., & Turner, A. (2006). *User-directed sentiment analysis: Visualizing the affective content of documents*. Proceedings of the Workshop on Sentiment and Subjectivity in Text, Sydney, Australia.

Greene, D., O'Callahan, D., & Cunningham, P. (2014). *How many topics? Stability analysis for topic models*. In T.Calders, F.Esposito, E.Hüllermeier, & R.Meo (Eds.), *Machine learning and knowledge discovery in databases* (Vol. 87352, pp. 498–513). Berlin, Germany: Springer.

Grimmelmann, J. (2015, May27). *Do you consent? If tech companies are going to experiment on us, they need better ethical oversight*. *Slate*. Retrieved from http://www.slate.com/articles/technology/future_tense/2015/05/facebook_emotion_contagion_study_tech_companies_need_irb_review.html

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in

Senate press releases. *Political Analysis*, 18(1), 1–35.

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

Günther, E., & Quandt, T. (2016). Word counts and topic models: Automated text analysis methods for digital journalism research. *Digital Journalism*, 4(1), 75–88.

Haigh, C., & Jones, N. (2005). An overview of the ethics of cyber-space research and the implications for nurse educators. *Nurse Education Today*, 25(1), 3–8.

Haigh, C., & Jones, N. (2007). *Techno-research and cyber ethics: Research using the Internet*. In T.Long & M.Johnson (Eds.), *Research ethics in the real world: Issues and solutions for health and social care* (pp. 157–174). Philadelphia, PA: Elsevier Health Sciences.

Hair, N., & Clark, M. (2007). *The ethical dilemmas and challenges of ethnographic research in electronic communities*. *International Journal of Market Research*, 49(6). Retrieved from https://www.mrs.org.uk/ijmr/archive#Articles

Hakimnia, R., Holmström, I., Carlsson, M., & Höglund, A. (2014). Exploring the communication between telenurse and caller—A critical discourse analysis. *International Journal of Qualitative Studies on Health and Well-Being*, 9, 1–9.

Halberstadt, A., Langley, H., Hussong, A., Rothenberg, W., Coffman, J., Mokrova, I., & Costanzo, P. (2016). Parents' understanding of gratitude in children: A thematic analysis. *Early Childhood Research Quarterly*, 36, 439–451.

Hanna, A. (2013). Computer-aided content analysis of digitally enabled movements. *Mobilization*, 18(4), 367–388.

Hardy, C. (2001). Researching organizational discourse. *International Studies of Management & Organization*, 31(3), 25–47.

Harris, J. (2011). *Word clouds considered harmful*. *Nieman Journalism Lab*. Retrieved June26, 2015, from http://www.niemanlab.org/2011/10/word-clouds-considered-harmful

Hatzivassiloglou, V., & McKeown, K. (1997). *Predicting the semantic orientation of adjectives*. Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (pp. 174–181). Stroudsburg, PA: Association for Computational Linguistics.

Hellsten, I., Dawson, J., & Leydesdorff, L. (2010). Implicit media frames: Automated analysis of public debate on artificial sweeteners. *Public Understanding of Science*, 19(5), 590–608.

Herrera, Y. M., & Braumoeller, B. F. (2004, Spring). *Symposium: Discourse and content analysis*. *Qualitative Methods Newsletter*, 15–19. Retrieved from http://www.braumoeller.info/wp-content/uploads/2012/12/Discourse-Content-Analysis.pdf

Hirschman, E. C. (1987). People as products: Analysis of a complex marketing exchange. *Journal of Marketing*, 51(1), 98–108.

Hardie, A., Koller, V., Rayson, P., & Semino, E. (2007). *Exploiting a semantic annotation tool for metaphor analysis*. In M.Davies, P.Rayson, S.Hunston, & P.Danielsson (Eds.), Proceedings of the Corpus Linguistics 2007 Conference. Retrieved June27, 2015, from ucrel.lancs.ac.uk/people/paul/publications/HardieEtAl_CL2007.pdf

Hart, C. (2010). *Critical discourse analysis and cognitive science: New perspectives on immigration discourse*. Basingstoke, England: Palgrave Macmillan.

Hayes, A., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77–89.

Heath, C., & Luff, P. (2000). *Technology in action*. Cambridge, England: Cambridge University Press.

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2), 135–175.

Henderson, S., & Segal, E. (2013). Visualizing qualitative data in evaluation research. *New Directions for Evaluation*, 139, 53–71.

Heritage, J., & Raymond, G. (2005). The terms of agreement: Indexing epistemic authority and subordination in talk-in-interaction. *Social Psychology Quarterly*, 68(1), 15–38.

Hewson, C. (2014). *Qualitative approaches in Internet-mediated research: Opportunities, issues, possibilities*. In P.Leavy (Ed.), *The Oxford handbook of qualitative research* (pp. 423–452). New York, NY: Oxford University Press.

Hewson, C., & Laurent, D. (2012). *Research design and tools for Internet research*. In J.Hughes (Ed.), *SAGE Internet research methods: Volume 1*. Thousand Oaks, CA: Sage.

Hewson, C., Vogel, C., & Laurent, D. (2015). *Internet research methods: A practical guide for the behavioural*

and social sciences. Thousand Oaks, CA: Sage.

Hewson, C., Yule, P., Laurent, D., & Vogel, C. (Eds.). (2003). *Internet research methods: A practical guide for the social and behavioural sciences*. Thousand Oaks, CA: Sage.

Hine, C. (2000). *Virtual ethnography*. Thousand Oaks, CA: Sage.

Hoffman, M. (1999). Problems with Peirce's concept of abduction. *Foundations of Science*, 4(3), 271–305.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage.

Holstein, J., & Gubrium, J. (2011). *Varieties of narrative analysis*. Thousand Oaks, CA: Sage.

Howell, K. (2013). *An introduction to the philosophy of methodology*. Thousand Oaks, CA: Sage.

Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 168–177). New York, NY: Association for Computing Machinery.

Hugo, R. (1992). *In defense of creative writing classes*. *The triggering town: Lectures and essays on poetry and writing* (pp. 53–66). New York, NY: W. W. Norton.

Ignatow, G. (2003). "Idea hjamsters" on the "bleeding edge": Profane metaphors in high technology argon. *Poetics*, 31(1), 1–22.

Ignatow, G. (2004). Speaking together, thinking together? Exploring metaphor and cognition in a shipyard union dispute. *Sociological Forum*, 19(3), 405–433.

Ignatow, G. (2009). Culture and embodied cognition: Moral discourses in Internet support groups for overeaters. *Social Forces*, 88(2), 643–669.

Ignatow, G., & Williams, A. T. (2011). New media and the "anchor baby" boom. *Journal of Computer-Mediated Communication*, 17(1), 60–76.

Ilieva, J., Baron, S., & Healey, N. M. (2002). Online surveys in marketing research: Pros and cons. *International Journal of Market Research*, 44(3), 361–376.

Illia, L., Sonpar, K., & Bauer, M. W. (2014). Applying co-occurrence text analysis with Alceste to studies of impression management. *British Journal of Management*, 25 (2), 352–372.

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1), 89–106.

James, W. (1975). *Pragmatism: A new name for some old ways of thinking*. Cambridge, MA: Harvard University Press. (Original work published 1907)

James, W. (1975). *The meaning of truth*. Cambridge, MA: Harvard University Press. (Original work published 1909)

Jockers, M. L. (2010, March19). *Who's your DH blog mate: Match-making the day of DH bloggers with topic modeling*. *Matthew L. Jockers*. Retrieved from http://www.matthewjockers.net/2010/03/19/whos-your-dh-blog-mate-match-making-the-day-of-dh-bloggers-with-topic-modeling

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41(6), 750–769.

Johnson-Laird, P. N. (1983). *Mental models: Toward a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Jones, M. V., Coviello, Y., & Tang, Y. K. (2011). International entrepreneurship research (1989–2009), A domain ontology and thematic analysis. *Journal of Business Venturing*, 26(6), 632–649.

Jurafsky, D., & Martin, J. (2009). *Speech and language processing*. Upper Saddle River, NJ: Prentice Hall.

Kallus, N. (2014). *Predicting crowd behavior with big public data*. *WWW '14 Companion Proceedings of the 23rd International Conference on World Wide Web*, 625–630. doi:

Kaplan, D. (Ed.). (2009). *Readings in the philosophy of technology*. Lanham, MD: Rowman & Littlefield.

Kassarjian, H. (1977). Content analysis in consumer research. *Journal of Consumer Research*, 4(1), 8–18.

Kim, S.-M., & Hovy, E. (2006). *Identifying and analyzing judgment opinions*. Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics.

King, A. (2008). *In vivo coding*. In L.Given (Ed.), *The SAGE encyclopedia of qualitative research methods*. Thousand Oaks, CA: Sage.

Klein, D., & Manning, C. D. (2004). *Corpus-based induction of synactic structure: Models of dependency and constituency*. *Proceedings of the Forty-Second Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.

Kleinman, D., & Moore, K. (2014). *Routledge handbook of science, technology, and society*. New York, NY: Routledge.

Koller, V., & Mautner, G. (2004). *Computer applications in critical discourse analysis*. *Applying English grammar* (pp. 216–228). London, England: Hodder and Stoughton.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 401–412.

Kovecses, Z. (2002). *Metaphor: A practical introduction*. Oxford, England: Oxford University Press.

Kozinets, R. V. (2002). The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research*, 39(1), 61–72.

Kozinets, R. V. (2009). *Netnography: Doing ethnographic research online*. Thousand Oaks, CA: Sage.

Kramer, A., Guillory, J., & Hancock, J. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788–8790.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Krishnamurthy, R. (1996). *Ethnic, racial and tribal: The language of racism?* In C. R.Caldas-Coulthard & M.Coulthard (Eds.), *Texts and practices: Readings in critical discourse analysis* (pp. 128–149). London, England: Routledge.

Krueger, R. A., & Casey, M. A. (2014). *Focus groups: A practical guide for applied research*. Thousand Oaks, CA: Sage.

Kuckartz, U. (2014). *Qualitative text analysis: A guide to methods, practice, and using software*. Thousand Oaks, CA: Sage.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W., & Waletzky, J. (1967). *Narrative analysis*. In J.Helm (Ed.), *Essays on the verbal and visual arts* (pp. 12–44). Seattle: University of Washington Press.

Lahlou, S. (1996). A method to extract social representations from linguistic corpora. *Japanese Journal of Experimental Social Psychology*, 35(3), 278–291.

Laird, E. A., McCance, T., McCormack, B., & Gribben, B. (2015). Patients' experiences of in-hospital care when nursing staff were engaged in a practice development programme to promote person-centredness: A

narrative analysis study. *International Journal of Nursing Studies*, 52(9), 1454–1462.

Lakoff, G. (1987). *Women, fire, and dangerous things. What categories reveal about the mind*. Chicago, IL: University of Chicago Press.

Lakoff, G. (1996). *Moral politics*. Chicago, IL: University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh*. New York, NY: Basic Books.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *Psychology of Learning and Motivation*, 41, 43–84.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3), 259–284.

Lasswell, H. (1927). Propaganda technique in the world war. *American Political Science Review*, 21(3), 627–631.

Lazard, A., Scheinfeld, E., Bernhardt, J., Wilcox, G., & Suran, M. (2015). Detecting themes of public concern: A text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. *American Journal of Infection Control*, 43(10), 1109–1111.

Lee, B., Riche, N. H., Karlson, A. K., & Carpendale, S. (2010). SparkClouds: Visualizing trends in tag clouds. *Visualization and Computer Graphics, IEEE Transactions on Knowledge and Data Engineering*, 16(6), 1182–1189.

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

Leondar-Wright, B. (2014). *Missing class: Strengthening social movement groups by seeing class cultures*. Ithaca, NY: Cornell University Press.

LeRoux, B., & Rouanet, H. (2010). *Multiple correspondence analysis*. Thousand Oaks, CA: Sage.

Lesk, M. (1986). *Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone*. Proceedings of the SIGDOC Conference 1986 (pp. 24–26). New York, NY: ACM.

Levenberg, A., Pulman, S., Moilanen, K., Simpson, E., & Roberts, S. (2014). *Predicting economic indicators*

from web text using sentiment composition. Retrieved from http://www.robots.ox.ac.uk/~parg/pubs/sentiment_ICICA2014.pdf

Levina, N., & Arriaga, M. (2012). Distinction and status production on user-generated content platforms: Using Bourdieu's theory of cultural production to understand social dynamics in online fields. *Information Systems Research*, 25(3), 468–488.

Levy, K., & Franklin, M. (2013). Driving regulation: Using topic models to examine political contention in the U.S. trucking industry. *Social Science Computer Review*, 32(2), 182–194.

Light, R., & Cunningham, J. (2016). Oracles of peace: Topic modeling, cultural opportunity, and the Nobel Peace Prize, 1902–2012. *Mobilization: An International Quarterly*, 21(1), 43–64.

Lindseth, A., & Norberg, A. (2004). A phenomenal hermeneutical method for researching lived experience. *Scandinavian Journal of Caring Sciences*, 18(2), 145–153.

Lipton, P. (2003). *Inference to the best explanation*. New York, NY: Routledge.

Liu, B., & Mihalcea, R. (2007). *Of men, women, and computers: Data-driven gender modeling for improved user interfaces*. Paper presented at the Proceedings of the International Conference on Weblogs and Social Media, Boulder, CO.

Lloyd, L., Kechagias, D., & Skiena, S. (2005). Lydia: A system for large-scale news analysis. *Processing and Information Retrieval*, 3372, 161–166.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). *Learning word vectors for sentiment analysis*. *Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA: Association for Computational Linguistics.

Macmillan, K. (2005). *More than just coding? Evaluating CAQDAS in a discourse analysis of news texts*. *Forum: Qualitative Social Research*, 6(3). Retrieved June27, 2015, from qualitative-research.net/index.php/fqs/article/view/28

Magnini, B., & Cavaglia, G. (2000). *Integrating subject field codes into WordNet*. Proceedings of the Conference on Language Resources and Evaluations (LREC-2000) (pp. 1413–1418). Athens, Greece.

Maguire, S., Hardy, C., & Lawrence, T. (2004). Institutional entrepreneurship in emerging fields: HIV/AIDS treatment advocacy in Canada. *Academy of Management Journal*, 47(5), 657–679.

Mairesse, F., Walker, M., Mehl, M., & Moore, R. (2007). Using linguistic cures for the automatic recognition of

personality in conversation and text. *Journal of Artificial Intelligence Research*, 30, 457–501.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313–330.

Marwick, B. (2013). *Discovery of emergent issues and controversies in anthropology using text mining, topic modeling, and social network analysis of microblog content*. In C.Yonghua & Y.Zhao (Eds.), *Data mining applications with R*, 63–93. Cambridge, England: Academic Press.

Mason, Z. J. (2004). Cormet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1), 23–44.

Mathews, A. S. (2005). Power/knowledge, Power/ignorance: Forest fires and the state in Mexico. *Human Ecology*, 33(6), 795–820.

McCallum, A., & Li, W. (2003). *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons*. Proceedings of the Seventh Conference on Natural Language Learning. Stroudsburg, PA: Association for Computational Linguistics.

McCallum, A., & Nigam, K. (1998). *A comparison of event models for Naive Bayes text classification*. Paper presented at the AAAI-98 Workshop on Learning for Text Categorization.

McFarland, D., Ramage, D., Chuang, J., Heer, J., Manning, C., & Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics*, 41(6), 607–625.

Merkl-Davies, D. M., & Koller, V. (2012). “Metaphoring” people out of this world: A critical discourse analysis of a chairman’s statement of a UK defence firm. *Accounting Forum*, 36(3), 178–193.

Merton, R. K. (1949). *On sociological theories of the middle range*. In R. K.Merton, *Social theory and social structure* (pp. 39–53). New York, NY: Free Press.

Meyer, M. (2014, June30). *Everything you need to know about Facebook’s controversial emotion experiment*. *Wired*. Retrieved from http://www.wired.com/2014/06/everything-you-need-to-know-about-facebooks-manipulative-experiment

Mihalcea, R. (2007). *Using Wikipedia for automatic word sense disambiguation*. *Proceedings of NAACL HLT* (pp. 196–203). Retrieved June27, 2015, from aclweb.org/anthology/N07-1025

Mihalcea, R., Banea, C., & Wiebe, J. (2007). *Learning multilingual subjective language via cross-lingual projections*. Paper presented at the Proceedings of the Association for Computational Linguistics, Prague,

Czech Republic.

Mihalcea, R., & Strapparava, C. (2009). *The lie detector: Explorations in the automatic recognition of deceptive language*. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 309–312). Stroudsburg, PA: Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. *Advances in neural information processing systems* (pp. 3111–3119).

Miles, M. B., & Huberman, A. M. (1994). *Data management and analysis methods*. Thousand Oaks, CA: Sage.

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

Mische, A. (2014). Measuring futures in action: Projective grammars in the Rio+20 debates. *Theory & Society*, 43(3–4), 437–464.

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

Moser, K. (2000). *Metaphor analysis in psychology—Method, theory, and fields of application*. *Forum: Qualitative Social Research*, 1(2), Art. 21. Retrieved from http://nbn-resolving.de/urn:nbn:de:0114-fqs0002212

Mukherjee, A., & Liu, B. (2012). *Aspect extraction through semi-supervised modeling*. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. (pp. 339–348).Stroudsburg, PA: Association for Computational Linguistics.

Mützel, S. (2015). Facing big data: Making sociology relevant. *Big Data & Society*, 2(2), 1–4.

Nakagawa, T., Inui, K., & Kurohashi, S. (2010). *Dependency tree-based sentiment classification using CRFs with hidden variables*. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 786–794). Stroudsburg, PA: Association for Computational Linguistics.

Narayanan, A., & Shmatikov, V. (2008). *Robust de-anonymization of large sparse datasets (How to break anonymity of the Netflix prize dataset)*. *IEEE Symposium on Security & Privacy*, Oakland, CA. Retrieved from http://arxiv.org/pdf/cs/0610105v2

Narayanan, A., & Shmatikov, V. (2009). *De-anonymizing social networks*. *IEEE Symposium on Security &*

Privacy, Oakland, CA. Retrieved from http://www.cs.utexas.edu/~shmat/shmat_oak09.pdf

Navigli, R., & Ponzetto, S. (2012). *Artificial Intelligence*, 193, 217–250.

Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Newton, H., & Frieder, O. (2013). Metaphor identification in large texts corpora. *PLOS ONE*, 8(4), 1–9.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5),665–675.

Noel-Jorand, M.-C., Reinert, M., Bonnon, M., & Therme, P. (1995). Discourse analysis and psychological adaptation to high altitude hypoxia. *Stress Medicine*, 11(1), 27–39.

O'Halloran, K., & Coffin, C. (2004). *Checking over-interpretation and under-interpretation: Help from corpora in critical linguistics*. *Text and Texture: Systemic Functional Viewpoints on the Nature and Structure of Text*, 275–297.

O'Keefe, A., & Walsh, S. (2012). Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education. *Corpus Linguistics and Linguistic Theory*, 8(1), 159–181.

Olthouse, J. (2014). How do preservice teachers conceptualize giftedness? A metaphor analysis. *Roeper Review*, 36(2), 122–132.

O'Mara-Shimek, M., Guillén-Parra, M., & Ortega-Larrea, A. (2015). Stop the bleeding or weather the storm? Crisis solution marketing and the ideological use of metaphor in online financial reporting of the stock market crash of 2008 at the New York Stock Exchange. *Discourse & Communication*, 9(1), 103–123.

Ortony, A., Clore, G. L., & Collins, A. (1990). *The cognitive structure of emotions*. New York, NY: Cambridge University Press.

Osmond, A. (2016). *Academic writing and grammar for students*. Thousand Oaks, CA: Sage.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). *Finding deceptive opinion spam by any stretch of the imagination*. *Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies—Volume 1 Association for Computational Linguistics* (pp. 309–319). Stroudsburg, PA: Association for Computational Linguistics.

Pang, B., & Lee, L. (2004). *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts*. *Proceedings of the Forty-Second Annual Meeting on Association for Computational Linguistics*. (pp. 271–278).Stroudsburg, PA: Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–35.

Parker, I. (1992). *Discourse dynamics: Critical analysis for social and individual psychology*. London, England: Routledge.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Thousand Oaks, CA: Sage.

Patton, M. Q. (2014). *Qualitative research & evaluation methods: Integrating theory and practice* (4th ed.). Thousand Oaks, CA: Sage.

Pauca, V. P., Shahnaz, F., Berry, M. W., & Plemmons, R. J. (2004). *Text mining using non-negative matrix factorizations*. Proceedings of the Fourth SIAM International Conference on Data Mining. Retrieved June27, 2015, from epubs.siam.org/doi/pdf/10.1137/1.9781611972740.45

Peirce, C. S. (1901). Truth and falsity and error. *Dictionary of Philosophy and Psychology*, 2, 716–720.

Pennebaker, J. W., Francis, M., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program*. Mahwah, NJ: Lawrence Erlbaum.

Pennebaker, J. W., & King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77, 1296–1312.

Phillips, N., & Hardy, C. (2002). *Discourse analysis: Investigating processes of social construction*. Thousand Oaks, CA: Sage.

Plummer, K. (1995). *Telling sexual stories: Power, change and social worlds*. London, England: Routledge.

Popping, R. (1997). *Computer programs for the analysis of texts and transcripts*. In *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 209–221). Mahwah, NJ: Lawrence Erlbaum.

Potter, J., & Wetherell, M. (1987). *Discourse and social psychology: Beyond attitudes and behavior*. Newbury Park, CA: Sage.

Propp, V. (1968). *Morphology of the folktale*. Austin: University of Texas Press.

Puschmann, C., & Burgess, J. (2014). Big data, big questions: Metaphors of big data. *International Journal of Communication*, 8, 1690–1709.

Quinlan, J. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1), 209–228.

Ratnaparkhi, A. (1996, May). *A maximum entropy model for part-of-speech tagging*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (Vol. 1., pp. 133–142).

Ravitch, S. M., & Riggan, J. M. (2016). *Reason & rigor: How conceptual frameworks guide research*. Thousand Oaks, CA: Sage.

Rees, C. E., Knight, L. V., & Wilkinson, C. E. (2007). Doctors being up there and we being down here: A metaphorical analysis of talk about student/doctor–patient relationships. *Social Science and Medicine*, 65(4), 725–737.

Resnik, P., Garron, A., & Resnik, R. (2013). *Using topic modeling to improve prediction of neuroticism and depression in college students*. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1348–1353). Stroudsburg, PA: Association for Computational Linguistics.

Richards, L., & Morse, J. (2013). *README FIRST for a user's guide to qualitative methods*. Thousand Oaks, CA: Sage.

Richardson, D. C., Spivey, M. J., Barsalou, L. W., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive Science*, 27(5), 767–780.

Ricoeur, P. (1991). Narrative identity. *Philosophy Today*, 35(1), 73–81.

Riloff, E., & Jones, R. (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 Innovative Applications of Artificial Intelligence (pp. 474–479). Menlo Park, CA: American Association for Artificial Intelligence.

Roberts, C. W. (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah, NJ: Lawrence Erlbaum.

Roberts, C. W. (2008). *The fifth modality: On languages that shape our motivations and cultures*. Leiden, Netherlands: Brill Publishers.

Roberts, C. W., Zuell, C., Landmann, J., & Wang, Y. (2010). Modality analysis: A semantic grammar for imputations of intentionality in texts. *Quality & Quanitity*, 44(2), 239–257.

Roberts, C. W., Popping, R., & Pan, Y. (2009). Modalities of democratic transformation forms of public

discourse in Hungary's latest newspaper, 1990–1997. *International Sociology*, 24(4), 498–525.

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

Roderburg, S. (1998). *Sprachliche konstruktion der wirklichkeit. Metaphern in therapiegesprächen*. Wiesbaden, Germany: Deutscher Universitäts Verlag.

Roget, P. (1987). *Roget's thesaurus of English words and phrases*. New York, NY: Longman. (Original work published 1911)

Rosenwald, G. C., & Ochberg, R. L. (1992). *Storied lives: The cultural politics of self-understanding*. New Haven, CT: Yale University Press.

Rousselière, D., & Vézina, M. (2009). Constructing the legitimacy of a financial cooperative in the cultural sector: A case study using textual analysis. *International Review of Sociology: Revue Internationale de Sociologie*, 19(2), 241–261.

Ruan, X., Wilson, S., & Mihalcea, R. (2016, August7–12). *Finding optimists and pessimists on Twitter*. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 320–325). Berlin, Germany.

Ruiz Ruiz, J. (2009). *Sociological discourse analysis: Methods and logic*. *Forum: Qualitative Social Research*, 10(2). Retrieved June27, 2015, from qualitative-research.net/index.php/fqs/article/view/1298/2882

Ryan, G. W., & Bernard, H. R. (2010). *Analyzing qualitative data: Systematic approaches*. Thousand Oaks, CA: Sage.

Sahpazia, P., & Balamoutsoua, S. (2015). Therapists' accounts of relationship breakup experiences: A narrative analysis. *European Journal of Psychotherapy & Counselling*, 17(3), 258–276.

Salganik, M. (in press). *Bit by bit: Social research in the digital age*. Retrieved from http://www.bitbybitbook.com

Salmons, J. (2014). *Qualitative online interviews*. Thousand Oaks, CA: Sage.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, PA: Addison-Wesley.

Santa Ana, O. (2002). *Brown tide rising metaphors of Latinos in contemporary American public discourse*. Austin: University of Texas Press.

Sapir, J., & Crocker, J. (Eds.). (1977). *The social use of metaphor: Essays on the anthropology of rhetoric*. Philadelphia: University of Pennsylvania Press.

Saussure de, F. (1959). *Course in general linguistics*. New York, NY: The Philosophical Library.

Schmidt, B. M. (2012). *Words alone: Dismantling topic models in the humanities*. *Journal of Digital Humanities*, *2*(1).

Schmitt, R. (2000). *Notes towards the analysis of metaphor*. *Forum Qualitative Social Research*, 1(1).

Schmitt, R. (2005). Systematic metaphor analysis as a method of qualitative research. *The Qualitative Report*, 10(2), 358–394.

Schonhardt-Bailey, C. (2013). *Deliberating American monetary policy: A textual analysis*. London, England: MIT Press.

Schuster, J., Beune, E., & Stronks, K. (2011). Metaphorical constructions of hypertension among three ethnic groups in the Netherlands. *Ethnicity and Health*, 16(6), 583–600.

Schwandt, T. A. (2001). *Dictionary of qualitative research*. Thousand Oaks, CA: Sage.

Shaw, C., & Nerlich, B. (2015). Metaphor as a mechanism of global climate change governance: A study of international policies, 1992–2012. *Ecological Economics*, 109, 34–40.

Shepherd, A., Sanders, C., Doyle, M., & Shaw, J. (2015). *Using social media for support and feedback by mental health service users: Thematic analysis of a Twitter conversation*. *BMC Psychiatry*, 15(29).

Silverman, D. (1993). *Interpreting qualitative data: Methods for analyzing talk, text and interaction*. Newbury Park, CA: Sage.

Silverman, D. (Ed.). (2016). *Qualitative research*. Thousand Oaks, CA: Sage.

SnowC. P. (2013). *The two cultures and the scientific revolution*. London, England: Martino Fine Books. (Original work published 1959)

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank*. Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Soroka, S., Stecula, D., & Wlezien, C. (2015). It's (change in) the (future) economy, stupid: Economic indicators, the media, and public opinion. *American Journal of Political Science*, 59(2), 457–474.

Speed, G. J. (1893). Do newspapers now give the news?*Forum*, 15, 705–711.

Spradley, J. P. (1972). *Adaptive strategies of urban nomads: The ethnoscience of tramp culture*. In T.Weaver & D. J.White (Eds.), *The anthropology of urban environments*. Boulder, CO: Society for Applied Anthropology.

Stark, A., Shafran, I., & Kaye, J. (2012). *Hello, who is calling?: Can words reveal the social nature of conversations?*Proceedings of the 2012 Conference of the North American Chapters of the Association for Computational Linguistics: Human Language Technologies (pp. 112–119).

Stone, P. J., Dunphry, D., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Stone, P. J., & Hunt, E. B. (1963). *A computer approach to content analysis: Studies using the General Inquirer system*. AFIPS ’63 (Spring) Proceedings of the May 21–23, 1963, Spring Joint Computer Conference (pp. 241–256). doi:

Strachan, J., Yellowlees, G., & Quigley, A. (2015). General practitioners’ assessment of, and treatment decisions regarding, common mental disorder in older adults: Thematic analysis of interview data. *Ageing and Society*, 35(1), 150–168.

Strapparava, C., & Mihalcea, R. (2007). *SemEval-2007 task 14: Affective text*. Proceedings of the Fourth International Workshop on the Semantic Evaluations, Prague, Czech Republic (pp. 70–74). Stroudsburg, PA: Association for Computational Linguistics.

Strapparava, C., & Valitutti, A. (2004). *WordNet-Affect: An affective extension of WordNet*. Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Strauss, C. (1992). *What makes Tony run? Schemas as motives reconsidered*. In R.D’Andrade & C.Strauss (Eds.), *Human motives and cultural models* (pp. 191–224). Cambridge, England: Cambridge University Press.

Stroet, K., Opdenakker, M.-C., & Minnaert, A. (2015). Need supportive teaching in practice: A narrative analysis in schools with contrasting educational approaches. *Social Psychology of Education*, 18(3), 585–613.

Strunk, W., & White, E. B. (1999). *The elements of style* (4th ed.). New York, NY: Pearson.

Stubbs, M. (1994). Grammar, text, and ideology: Computer-assisted methods in the linguistics of representation. *Applied Linguistics*, 15(2), 201–223.

Sudhahar, S., Franzosi, R., & Cristianini, N. (2011). Automating quantitative narrative analysis of news data.

JMLR: Workshop and Conference Proceedings, 17, 63–71.

Sudweeks, F., & Rafaeli, S. (1996). *How do you get a hundred strangers to agree: Computer mediated communication and collaboration*. In T. M.Harrison & T. D.Stephen (Eds.), *Computer networking and scholarship in the 21st century university* (pp. 115–136). New York, NY: SUNY Press.

Sun, Y., & Jiang, J. (2014). Metaphor use in Chinese and US corporate mission statements: A cognitive sociolinguistic analysis. *English for Specific Purposes*, 33, 4–14.

Sveningsson, M. (2003). *Ethics in Internet ethnography*. *International Journal of Global Information Management*, 11(3). Retrieved from http://www.irma-international.org/viewtitle/28292

Sweeney, L. (2003). *Navigating computer science research through waves of privacy concerns: Discussions among computer scientists at Carnegie Mellon University*. *Tech Report*, CMU CS 03-165, CMU-ISRI-03-102. Pittsburgh, PA.

Sweetser, E. (1990). *From etymology to pragmatics: The mind-body metaphor in semantic structure and semantic change*. Cambridge, England: Cambridge University Press.

Sword, H. (2012a). *Stylish academic writing*. Cambridge, MA: Harvard University Press.

Sword, H. (2012b, July23). *Zombie nouns*. *New York Times*.

Takamura, H., Inui, T., & Okumura, M. (2006). *Latent variable models for semantic orientations of phrases*. *Proceedings of the Eleventh Meeting of the European Chapter of the Association for Computational Linguistics* (pp. 201–208). Trento, Italy.

Tashakkori, A. M., & Teddlie, C. B. (2010). *SAGE handbook of mixed methods in social & behavioral research* (2nd ed.). Thousand Oaks, CA: Sage.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54.

Teddlie, C. B., & Tashakkori, A. M. (Eds.). (2008). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: Sage.

Toerien, M., & Wilkinson, S. (2004). Exploring the depilation norm: A qualitative questionnaire study of women's body hair removal. *Qualitative Research in Psychology*, 1(1), 69–92.

Toor, R. (2012, July2). *Becoming a "stylish" writer*. *Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Becoming-a-Stylish-Writer/132677

Törnberg, A., & Törnberg, P. (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society*, 27(4), 401–422.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003, May). *Feature-rich part-of-speech tagging with a cyclic dependency network*. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Technology—Volume 1 (pp. 173–180). Stroudsburg, PA: Association for Computational Linguistics.

Trappey, C., Wu, H., Liu, K., & Lin, F. (2013, September11–13). *Knowledge discovery of service satisfaction based on text analysis of critical incident dialogues and clustering methods*. 2013 IEEE 10th International Conference on e-Business Engineering (pp. 265–270). Coventry, United Kingdom: ICEBE2013.

Trochim, W. M. K. (1989). Concept mapping: Soft science or hard art?*Science Direct*, 12(1), 87–110.

Trochim, W. M. K., Cook, J. A., & Setze, R. (1994). Using concept mapping to develop a conceptual framework of staff's views of a supported employment program for individuals with severe mental illness. *Journal of Consulting and Clinical Psychology*, 62(4), 766–775.

Turney, P. D. (2001). *Mining the web for synonyms: PMI-IR versus LSA on TOEFL*. Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001) (pp. 491–502). Freiburg, Germany. NRC 44893.

Turney, P. D. (2002). *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. *Proceedings of the Fortieth Annual Meeting on Association for Computational Linguistics* (pp. 417–424). Stroudsburg, PA: Association for Computational Linguistics.

Turney, P. D., Neuman, Y., Assaf, D., & Cohen, Y. (2011). *Literal and metaphorical sense identification through concrete and abstract context*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 680–690). Stroudsburg, PA: Association for Computational Linguistics.

Uprichard, E. (2012). Describing description (and keeping causality): The case of academic articles on food and eating. *Sociology*, 47(2), 368–382.

Van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4(2), 249–283.

van Ham, F., Wattenberg, M., & Viégas, F. (2009). *Mapping text with phrase nets*. *IEEE Transactions on Visualization and Computer Graphics*, 15(6). Retrieved from http://ieeexplore.ieee.org/abstract/document/5290726

Van Herzele, A. (2006). A forest for each city and town: Story lines in the policy debate for urban forests in Flanders. *Urban Studies*, 43(3), 673–696. doi:

van Meter, K. M., & de Saint Léger, M. (2014). American, French & German sociologies compared through link analysis of conference abstracts. *Bulletin of Sociological Methodology*, 122(1), 26–45.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York, NY: Springer.

Viégas, F. B., & Wattenberg, M. (2008). TIMELINES: Tag clouds and the case for vernacular visualization. *Interactions*, 15(4), 49–52. doi:

Walejko, G. (2009). *Online survey: Instant publication, instant mistake, all of the above*. In E.Hargittai (Ed.), *Research confidential: Solutions to problems most social scientists pretend they never have* (pp. 101–115). Ann Arbor: University of Michigan Press.

Watson, M., Jones, D., & Burns, L. (2007). Internet research and informed consent: An ethical model for using archived emails. *International Journal of Therapy & Rehabilitation*, 14(9), 396–403.

Weale, A.Bicquelet, A., & Bara, J. (2012). Debating abortion, deliberative reciprocity and parliamentary advocacy. *Political Studies*, 60(3), 643–667.

Weisgerber, C., & Butler, S. H. (2009). *Visualizing the future of interaction studies: Data visualization applications as a research, pedagogical, and presentational tool for interaction scholars*. *The Electronic Journal of Communication*, 19(1–2). Retrieved June26, 2015, from http://www.cios.org/ejcpublic/019/1/019125.HTML

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.

Wetherell, M., & Edley, N. (1999). Negotiating hegemonic masculinity: Imaginary positions and psycho-discursive practices. *Feminism and Psychology*, 9(3), 335–356.

Wheeldon, J., & Ahlberg, M. (2012). *Visualizing social science research: Maps, methods, & meaning*. Thousand Oaks, CA: Sage.

Wheeldon, J., & Faubert, J. (2009). Framing experience: Concept maps, mind maps, and data collection in qualitative research. *International Journal of Qualitative Methods*, 8(3), 68–83.

White, H. (1978). *Tropics of discourse: Essays in cultural criticism*. Baltimore, MD: Johns Hopkins University Press.

White, P. W. (1924). *Quarter century survey of press content shows demand for facts*. *Editor and Publisher*, 57.

Wiebe, J., Bruce, R., & O'Hara, T. (1999). *Development and use of a gold-standard data set for subjectivity*

classifications. *Proceedings of the Thirty-Seventh Annual Meeting of the Association for Computational Linguistics* (pp. 246–253). Stroudsburg, PA: Association for Computational Linguistics.

Wiebe, J., & Mihalcea, R. (2006). *Word sense and subjectivity*. Paper presented at the Fourty-Fourth Annual Meeting of the Association for Computational Linguistics, Sydney, Australia.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3), 165–210.

Wilcox, D. F. (1900). The American newspaper: A study in social psychology. *The ANNALS of the American Academy of Political and Social Science*, 16(1), 56–92. doi:

Wilkinson, L., & Friendly, M. (2009). The history of the cluster heat map. *The American Statistician*, 63(9), 179–184.

Wilson, T. (2008). *Fine-grained subjectivity and sentiment analysis: Recognizing the intensity, polarity, and attitudes of private states* (PhD thesis, University of Pittsburgh).

Windelband, W. (1998). On history and natural science. *History and Theory*, 19, 165–185. (Original work published 1894)

Windelband, W. (2001). *A history of philosophy*. Cresskill, NJ: The Paper Tiger. (Original work published 1901)

Winkel, G. (2012). Foucault in the forests—A review of the use of “Foucauldian” concepts in forest policy analysis. *Forest Policy and Economics*, 16, 81–92.

Wofford, T. (2014, July28). *OkCupid co-founder: “We experiment on human beings . . . that’s how websites work.”Newsweek*. Retrieved from http://www.newsweek.com/okcupid-founder-we-experiment-human-beingsthats-how-websites-work-261741

Woodwell, D. (2014). *Research foundations: How do we know what we know?*Thousand Oaks, CA: Sage.

Wu, H., Liu, K., & Trappey, C. (2014). *Understanding customers using Facebook pages: Data mining users feedback using text analysis*. *IEEE*, 346–350.

Yarowsky, D. (2000). Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, 34(1), 179–186.

Yu, H., & Hatzivassiloglou, V. (2003). *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*. Paper presented at the Conference on Empirical Methods in

Natural Language Processing, Sapporo, Japan.

Yun, G. W., & Trumbo, C. W. (2000). *Comparative response to a survey executed by post, e-mail, & web form*. *Journal of Computer-Mediated Communication*, 6(1).

Zagibalov, T., & Carroll, J. (2008). *Automatic seed word selection for unsupervised sentiment classification of Chinese text*. Proceedings of the Twenty-Second International Conference on Computational Linguistics (pp. 1073–1080). Stroudsburg, PA: Association for Computational Linguistics.