Quest for Sociology
Revisiting Prevailing Understandings of a Discipline
with Computational Text Analyses of Dissertations

Josef Ginnerskov
Dissertation presented at Uppsala University to be publicly examined in Humanistiska teatern, Engelska parken, Thunbergsvägen 3C, Uppsala, Friday, 8 March 2024 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Sebastian Kohl (Freie Universität Berlin).

Abstract

What is sociology? For centuries sociologists have struggled to answer this question and repeatedly proclaimed that their discipline is in crisis. The problem has generated a field of its own, the sociology of sociology, where sociologists of knowledge offer concepts for how the paradigmatic status of discipline and its crisis ought to be understood. Yet, the foundation of these understandings has often been limited to conceptual reasonings, historical exposes, and anecdotes from prominent scholars. Following the increasing availability of digitized texts and the development of computational techniques, new venues have been opened for investigating the empirical bearing of what sociology is. This dissertation offers a synthesis of, and a contribution to, this growing literature at the intersection of the sociology of knowledge and computational social science.

The starting point is a review of literature in the sociology of sociology that has found that our discipline is believed to exist in a state of fragmentation, lacks a paradigm, and is conditioned by the context of its production. Akin to the supposed crisis, these conceptualizations are often taken for granted rather than being empirically put to test. This is why this dissertation aims to shed new light on the crisis of sociology by empirically scrutinizing prevailing disciplinary understandings with an interpretative and theory-driven methodological approach to computational text analysis (i.e., word correlation networks, topic modeling, stylometry, and shallow neural networks).

To account for textual representations of sociological knowledge that are firmly institutionalized and exist across different local contexts, hundreds of dissertations in this discipline published in Sweden between 1980 and 2019 by five main universities have been digitized to form two corpora – 380 full-texts and 850 abstracts. Using these corpora, the conceptualizations are operationalized to be able to scrutinize, and trace, reoccurring instances where dissertations allude to certain images of sociology, which, drawing on the work of Margaret Masterman, can be regarded as crude replicas of paradigms. The study design allows us to problematize prevailing understandings of what sociology is.

In contrast to the notion of fragmentation, the corpora are constituted by a core conditioned by local institutions attuned to different paradigmatic images of sociology. A discrepancy is also found between the two corpora where the abstracts appear to follow a divide between qualitative and quantitative research, and the full-texts are characterized by five paradigms with distinct methodological, epistemological, and ontological positions. These results suggest that the coexistence of multiple paradigms has been conflated with fragmentation and that sociologists tend to present their knowledge along the lines of simplified dichotomies. In response to the crisis, a more fruitful approach might be to embrace paradigm pluralism.

As a contribution to the sociology of knowledge, this dissertation is an example of how the methodological divide can be overcome by merging insights from the conceptual strand with a hermeneutical take on computational methods to empirically explore taken-for-granted assumptions behind the production of disciplinary knowledge.

Keywords: sociology of sociology, sociology of knowledge, sociological theory, paradigm theory, computational social science, digital humanities, text mining, dissertations

Josef Ginnerskov, Department of Sociology, Box 624, Uppsala University, SE-75126 Uppsala, Sweden.

© Josef Ginnerskov 2024

ISSN 0585-5551
ISBN 978-91-513-2016-8
URN urn:nbn:se:uu:diva-521096 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-521096)
Til Mette og Aske
Contents

1  Introduction ........................................................................................................ 9
   Aim ................................................................................................................. 17
   Outline of the Dissertation ........................................................................... 18

2  Reviewing Recent Trends in the Sociology of Sociology ...................... 21
   Textual Representations of Sociological Knowledge ..................................... 24
   Methods for Analyzing the Textual Representations ..................................... 35
   Theoretical Models of Sociological Knowledge .......................................... 43
   Towards a Computational Study of Dissertations ...................................... 55

3  Sociological Knowledge in Theory ......................................................... 59
   Paradigm Theories of Sociology .................................................................. 60
   Tracing Paradigms with Computers ............................................................. 66
   Delineating Crude Analogies through Dichotomies .................................... 71
   Ontological Dichotomies ............................................................................. 74
   Epistemological Dichotomies ...................................................................... 79
   Methodological Dichotomies ...................................................................... 83
   Putting the Theoretical Framework into Practice ....................................... 88

4  Methodology ............................................................................................ 91
   The Computational Imagination ................................................................. 92
   Interpreters Operating Machines ................................................................. 96
   Discovering and Measuring Patterns in Text Data ..................................... 100
   Sampling Dissertations and Generating Corpora ..................................... 104

5  Research Questions .................................................................................. 110

6  Word Constellations ................................................................................ 114
   Word Constellations on the Global Level ................................................. 115
   Communities of Word Constellations ....................................................... 122
   Community Allocations over Time and Place .......................................... 128
   Word Constellations in Time and Place Corpora ...................................... 135
   Closing Thoughts on Word Constellations .............................................. 144

7  Thematic Arrangements .......................................................................... 148
   Interpreting and Labeling Topics .............................................................. 150
   Topic Correlations and Communities ..................................................... 160
The Effect of Time on the Topic Proportions ........................................165
The Effect of Place on the Topic Proportions .......................................174
Closing Thoughts on Thematic Arrangements ......................................184

8 Linguistic Styles .................................................................................187
Discovering Styles with Bootstrapped Consensus Trees ......................190
Labeling Styles by Contrasting Z-scores .............................................195
Differentiating Styles with Close Reading ..........................................201
Spatial and Temporal Distributions of Styles ......................................209
Closing Thoughts on Linguistic Styles ................................................220

9 Distributional Meanings .....................................................................225
Training and Analyzing Word Embeddings .........................................227
Distributional Meanings of Sociology ..................................................228
Distributional Meanings of Society .......................................................241
Distributional Meaning of Dichotomies ..............................................252
Closing Thoughts on Distributional Meanings .....................................261

10 Concluding Discussion ....................................................................268
Questioning the Fragmentation of Sociological Knowledge ................272
Two Paradigmatic Images of Sociology ...............................................275
The Local Conditioning of Sociological Knowledge ............................279
Paradigm Pluralism – An Opening Out of the Crisis? ...........................285
Contributions and Future Work ............................................................289

Acknowledgements ..............................................................................295
References ............................................................................................299

Appendix A: Dissertations in Sociology Defended in Sweden ...............317
Appendix B: The Two Corpora and Their Limitations ............................323
Appendix C: Topic Modeling .................................................................328
Appendix D: Stylometry .........................................................................332
Appendix E: Word Embeddings ............................................................335
1 Introduction

Imagine that you were asked to prepare a lecture for a group of undergraduate students on the topic of *What is sociology?* Realizing that this seemingly easy question may not be as straightforward as it first seems, you open a textbook, or, perhaps more likely these days, an internet browser, and find a neat definition like *Sociology is the scientific study of society and social behavior* or *Sociology is the study of human social relationships and institutions.* These definitions, however, are only temporarily satisfactory since they call for problematic follow-up questions in the style of *But, what is society, then?* or *What distinguishes something ‘social’ from something ‘cultural’?* You also realize that a precocious student would figure out that human existence depends on society and, thus, could ask *Can ‘what is social’ really be distinguished from ‘what is human’?* You recognize that if you answer *Not really* to this question, you have ultimately accepted that *Sociology is the study of everything human.* This might be true but surely there must be a better way to explain the discipline, so you turn to the older neighboring sciences of society for advice. In contrast, these social sciences appear easier to define since they have, in some way, narrowed down their overarching research objects to specific domains of the social sphere like governance for political science or to a favored methodology (for example ethnography for anthropology). All of a sudden you feel a bit envious of economics with its mathematical and neoclassical models for studying the economy. Since sociology does not immediately appear to have a corresponding designated sphere of society or preferred methodology, you end up presenting the students with one of the vague definitions of sociology and strive to ensure that your classroom will be a space ‘liberated’ from any serious engagements with the questions of *What is sociology?*

Perhaps you are rooted in the discipline but find the narration above to be unrelatable or pompous – that this kind of ‘soul-searching’ is not only a non-issue for you as an individual sociologist but also for sociology in general. While the former position is a valid personal choice, the latter is, however, an empirical fallacy. Indeed, the problem of sociology can be found from the formative years of the discipline and their attempts to delineate the, at the time, ‘new science of society’ (e.g., Alatas, 2022) until today, which can be sensed in book titles like *The Crisis in Sociology* (Boudon, 1980), *Crisis in
Sociology (Lopreato, 1999), and Sociology After the Crisis (Lemert, 2004), to name but a few. At the turn of the 21st century, mainstream journals of the discipline have continued to publish articles on the problem of ‘finding’ sociology (e.g., Ballantine et al., 2016; House, 2019; Woolcock and Kim, 2000) and even full special issues have been dedicated to, for instance, “Bringing Sociology Home” (Meer et al., 2016, p. 838) or gaining insight by letting ‘blind non-sociologists describe the elephant of sociology’ (Gerholm and Gerholm, 1987). In many of these disciplinary outlets, commenters air various gloomy predictions, which can be illustrated with lines like “the current course of sociology will lead to academic self-destruction” (Lopreato 1999, p. xi) and “sociology has to be achieved against an internal tendency to self-subversion” (Holmwood, 2010, p. 650).

The most noticeable sociologists scrutinizing the disciplinary crisis can be found within the research strand referred to as the sociology of knowledge. Put in one sentence by a key figure in the field’s later developments, David Bloor (2004, p. 919), the sociology of knowledge “furnishes a theory of knowledge which exhibits knowing as a social process, and knowledge as a collective accomplishment”. Central to it are questions of what types of knowledge become generated, how it is disseminated in different societies, and how social conditions and cultural contexts shape its production (cf. Mulckay, 2016). The field is known to be as old as sociology itself and shares the trait that it is not easily overviewed, something which Robert K. Merton (1973, p. 10) noted in 1945: “To outline even the main currents of the sociology of knowledge in brief compass is to present none adequately and to do violence to all.” Indeed, the first half of the 20th century witnessed contributions spanning from grand analyses of knowledge within whole nations (e.g., Émile Durkheim and Karl Mannheim) to phenomenological accounts of how human knowledge is shaped by shared values and intersubjectivity (e.g., Max Scheler and Alfred Schütz). A continuation of the latter stream can be found in Peter L. Berger’s and Thomas Luckmann’s 1966 Treatise in the Sociology of Knowledge, one of the top-five most influential books of the 20th century according to the discipline representative,1 with its focus on the social construction of knowledge in everyday life. In the half-century succeeding this book, the sociology of knowledge has moved “from an examination of the contents of knowledge to the investigation of forms and practices of knowing” (cf. Swidler and Arditi, 1994, p. 306), primarily based on micro-level case studies.

As for empirical studies of knowledge generated in the sciences, Merton’s (1973, 1972) functionalist delineation of the sociology of science around the

---

middle of the 20th century – an active attempt to break from the traditional sociology of knowledge – and the so-called laboratory studies of the 1980s (see Knorr-Cetina, 1981; Latour and Woolgar, 1979) spearheaded the development of the productive sub-field known as the sociology of scientific knowledge or SSK (cf. Collins, 1983). In line with the overarching genealogy of the sociology of knowledge, it primarily focused on ethnographic accounts of how knowledge is generated and legitimized rather than paying any particular interest in the macro level, like Mannheim, or the content of the knowledge being studied (cf. Pels, 1996). Along with the overarching sociology of science, the sub-field of sociology of scientific knowledge has, over the last decades, more or less succumbed to the much larger and interdisciplinary field of science and technology studies or STS (cf. Shapin, 1995). The overarching shift in the sociology of knowledge towards SSK and STS can be read as exchanging questions of what (content) for questions of how (practice). Following this shift, the answer given to the question What is sociology? would be: Sociology is whatever sociologists do.

However, since sociology presumably includes a span from the most humanist and qualitative to the most science-embracing and quantitative of social scientists, a micro-study of sociologists in situ would hardly be able to capture this diversity.2 Fortunately, a deviant development unfolds when turning to studies of sociological knowledge, which for the last century have maintained a focus on conceptual and historical studies (cf. Remmling, 2022). It is also, naturally, in this part of the sociology of knowledge where the crisis of sociology is given the most attention. While traces of this narrative can be followed back to the era of classical sociology at the turn of the 19th century, proper theories – of sociology in general and its crisis in particular – caught wind with what Robert K. Merton (1972, p. 9) referred to as “some prime examples” of “the application of the sociology of knowledge to the special case of sociology itself”, namely: The Coming Crisis of Western Sociology (Gouldner, 1970), A Sociology of Sociology (Friedrichs, 1970), and The Phenomenon of Sociology (Tiryakian, 1971).

Since the early 1970s, a series of studies focused on figuring out the discipline’s problem can be found under the loose heading of the sociology of sociology (e.g., Guggenheim, 2015; Jasper and Young, 2007; Mesny, 2009; Morin, 1985; Sokolov, 2019a; Wacquant, 1989). This literature contains a wide range of explanations for sociology’s current manifestation, such as that it is stuck in the ivory tower of professional sociology when it should be

---

2 In fact, one of the initial studies leading up to this dissertation, which was quickly abandoned due to trivial results, was an ethnography of how sociologists generate knowledge in their everyday lives at the department.
generating public sociology (Burawoy, 2005), or that it should abandon exegetical theory to develop explanatory theory (Sztomka, 2004). Nevertheless, a recurring framing of the crisis of sociology is the problem of fragmentation, where the knowledge that the discipline generates is spread out in a multitude of isolated sub-fields that lack any form of unifying principle and become increasingly autonomous vis-à-vis general sociology (e.g., O’Neill and Turner, n.d.; Steinmetz and Chae, 2002; Walby, 2021). While accepting this notion, a group of sociologists of knowledge has detached the assumed relationship between the character of sociological knowledge and the crisis of sociology, stating that the “fragmentation of sociology as its specialisms become its others is built into the very nature of sociology as a mode of intellectual investigation” (Scott, 2005, p. 3). Here, the fragmentation of sociological knowledge is turned into something of a virtue by envisioning it in the light of pluralism, an epistemological position stating that “it is possible to construct several different theories for the explanation of the same data”, and that doing so may contribute to the growth of knowledge” (Klima, 1972, p. 69).³

Fragmentation is not only debated in terms of specialisms but can, for instance, be found on the level of methodology, where sociology has been accused of suffering from methodological relativism (‘anything goes and there are too many methods’), segregation into incommensurable methodological camps, or being stuck with ‘outdated’ data collection methods such as interviews and surveys (cf. Savage and Burrows, 2007). Again, some sociologists of knowledge have advocated “for methodological pluralism and pragmatism and against methodological tribalism” (Lamont and Swidler, 2014, p. 153). However, this optimism in terms of methodology has been met with its share of skepticism, as some have argued that while the discipline appears to be ‘pluralistic’ in terms of accepting a wide share of methods, fewer sociologists are ‘pluralists’ in terms of actively applying different methods (cf. Payne et al., 2004). Another popular and long-lasting notion is that the work of the classics (or even pre-classics) can be summoned to find the ‘essence’ of sociology and the ‘remedy’ to its problems (e.g., Boudon, 2002; de Sousa Santos, 2009; Reed and Alexander, 2009; Walsh, 2013; Wernick, 2000). While the role the classics have played in sociology’s self-conception and teaching sociology is hard to ignore (cf. Ginniskov, 2023, 2021) – even if the so-called canon of classics is not free from biases (e.g., Bhambra, 2016; Connell, 1997) – several sociologists have questioned the value of the classics for conducting sociological research in contemporary times (e.g., Merton, 1967; Watts, 2014a).

³ Italics and quotation marks are found in the original quote.
When reading these theoretically sophisticated attempts to construe sociology within the sociology of knowledge, it is evident that there is one particular theory of science that guided the pioneering works of the 1970s (i.e., Bell, 1979; Eckberg and Hill, 1979; Eisenstadt, 1976; Friedrichs, 1970; Gouldner, 1970; Ritzer, 1975; Sztompka, 1979; Tiryakian, 1971) and that continues to inform the debates that take place well into the 21st century (e.g., Arluke, 2002; Brante, 1980, 1981; Eriksson, 1997, 1981; MacLean and Williams, 2008; Prandini, 2015; Steinmetz, 2004; Steinmetz and Chae, 2002; Wadsworth, 2005), namely Thomas S. Kuhn’s (1962) The Structure of Scientific Revolutions. Kuhn’s work is commonly referred to as paradigm theory and is founded on the idea that normal science rests upon a paradigm, a set of principles that all scientists in the field must agree upon that, so to speak, holds the science together. Hence, for real change (revolution!) to happen, the paradigm must be overthrown and exchanged for another paradigm, which is incommensurable with the first. Following Kuhn (1962: x, 37, 76), paradigms are “universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners” that supply scientists with “conceptual and instrumental tools”, and “so long as the tools of a paradigm supplies continue to prove capable of solving the problems it defines, science moves fastest and penetrates most deeply”.

Kuhn’s influence is evident in the most popular books of the first wave of the sociology of sociology debate hereby alluded to, speaking in somewhat mythical terms about the rise and what they, at that time, saw as the current fall of a ‘lost paradigm’ in sociology represented by the system-functionalism program associated with the lifework of Talcott Parsons. Indeed, in Alvin W. Gouldner’s (1970, p. 162) own words, “a single, organizing, intellectual center for the sociological community” arose from “Parsons’ system [which] was often a paradigm that gave coherence to the sociological community … [that] has been not so much exploded as picked apart and now is slowly expiring under the growing apathy of its audience”. In this narrative, the gradual disassembly of a paradigm founded on positivistic ideals is the root of the fragmentation process of sociological knowledge. Yet, regardless of its cause, the unsettling experience of living in a crisis lingers on in the sociological community, and paradigm theory is the lens through which it is comprehended. Through judging their seemingly sprawling discipline by the standard that a successful science rests upon a paradigm, a sense of order is instilled in the chaos.

Pierre Bourdieu (1993, p. 50) once wrote, “… one needs to ask what interest people have in doing the sociology of sociology”. Reading through the works above, it is often easier to find the agenda of how the various authors wish that sociology would look theoretically, than concrete evidence for how
sociology manifests itself empirically. From an analytical standpoint, the sociology of sociology project of the 20th century and its reflections on paradigms can be read as the more esoteric chapters in the general lore of the discipline, which, like all other stories, has its share of mythical heroes and villains. However, following the advice given in the book *What is Sociology?* by Norbert Elias (1978, p. 46), the role of the sociologist is to be “a hunter of myths” and sociology itself cannot be discriminated from this hunt. For the reader finding this statement desecrating, the prevailing understandings of sociology will be handled with great care and not be rejected by default since it is plausible that some of its elements hold the test of empirical scrutiny.

This dissertation builds on an empirical investigation of the crisis of sociology, as presented in this introduction, and, thus, makes a contribution to the sociology of knowledge, with special attention given to its subfield commonly referred to as the sociology of sociology. This field is in focus since it is where sociologists have made the greatest effort to theorize about the cause of the experiences of unease expressed by sociologists in discussions of their discipline. It is further argued that paradigm theory forms a foundation for the ways that influential voices within the discipline have turned the question of *what sociology is* into *what sociology is not* by questioning the scientific status and legitimacy of sociological knowledge. Central questions in that debate include *What are the characteristics of sociological knowledge and the way it progresses?* (e.g., Abbott, 2001; Bell, 1979), *Is sociology a single, dual, or multiple-paradigm science?* (e.g., Brante, 1981; Ritzer, 1975), and last but not least *Is paradigm theory even applicable to sociology (or is it too premature of a science)?* (e.g., Eckberg and Hill, 1979; Eriksson, 1997). As will be shown, theoretical works on these problems come in abundance but the empirical evidence is surprisingly sparse. Perhaps this is the reason why no major progress in understanding *what sociological knowledge is (not) has occurred during the last decades.*

As stated, sociologists of knowledge do not refrain from theorizing about the nature and current state of their discipline.4 However, these theories are, in general, based on a combination of conceptual and historical analysis (e.g., Abbott, 2001; Gouldner, 1970). Some of these studies lean more heavily on the philosophical method to find universal features of sociological knowledge, often based on works by classics in the larger countries of Europe and Northern America (e.g., Barnes, 2008; Bell, 1979; Steinmetz, 2004). In contrast, others address sociology’s particularities, essentially, by focusing on its historical developments in a single country (e.g., Larsson and

---

4 This is also most likely true for other social scientists even if their debates about their respective disciplines lie outside of the scope of this dissertation.
Magdalenić, 2015) or a few countries (e.g., Lepenies, 1988) – or, in rare cases, even a single department (e.g., Abbott, 1999) – holding that “each new emerging social context, in all its uniqueness, enlarges sociology’s subject” (Peel, 1972, p. 264). Besides this main bulk of studies in the style of intellectual history, the 20th century witnessed a set of empirical studies applying both qualitative and quantitative methods (e.g., Boalt, 1969; Lindholm-Romantschuk and Warner, 1996; Platt, 1995; Schneider and Hacker, 1973). While the 21st century has come with technological developments that were probably almost unthinkable at that time, sociologists of knowledge still today scrutinize sociology by applying the traditional tools of the discipline such as interviews (e.g., Hokka, 2019), archival/historical work (e.g., Halsey, 2004), surveys/questionaries (e.g., Wagenaar, 2004), and hand-coding text (e.g., Erola et al., 2015). This dissertation addresses this knowledge gap by applying comparatively novel computational tools, which are currently rather uncommon in sociology, in general, and the sociology of knowledge, in particular.

The accessibility of digital data and advances in computer-assisted techniques of the last century have attracted researchers in disciplines spanning from the natural sciences to the humanities to find new perspectives on questions of the same amplitude as the one this chapter began with (cf. Edelmann et al., 2020). Social scientists have been attuned to different terms to label their practice depending on their academic affiliation – such as culturomics (e.g., Michel et al., 2011) or digital sociology (e.g., Lupton, 2015; Selwyn, 2019) – yet, the most widespread and comprehending heading appears to be computational social science or CSS (cf. Lazer et al., 2009). While initially “being ignored” by social scientists (Watts, 2014b, p. 24), CSS is now a field housing interdisciplinary research societies, specialized journals, and even graduate programs. Not surprisingly, the stakes of promises the field is set to fulfil are also high. Through its “capabilities to collect and analyze data with an unprecedented breadth, depth, and scale”, CSS has been described as a force entailing “a paradigm shift in scientific research methods” (Chang et al., 2014, p. 68). Correspondingly, sociologists within the field have stated that “CSS has the potential to accomplish for sociology what the introduction of econometrics did for economics in the past half-century, i.e., to provide the relevant analytical tools and data needed to rigorously address the core questions of the discipline” (Keuschnigg et al., 2018, p. 8). There are, of course, other sociologists who have been less enthusiastic. Some of these voices have addressed that computational methods risk causing a naturalization of social phenomena (e.g., Törnberg and Törnberg, 2018). This dissertation will draw on precisely such tools for scrutinizing sociological knowledge, yet it seeks to take an agnostic view of the potential and limitations of CSS methods. This position will be developed along with the un-
folding of the chapters following this introduction but is particularly stressed in the methodology chapter.

For the problem sketched out in this introductory chapter, which at this juncture can be narrowed down to being about what sociological knowledge is (not) through the framework of paradigm theory, a computational analysis of the knowledge sociologists produce, which almost exclusively comes in the form of texts, seems particularly promising. One of the principal strands of CSS deals with computer-guided text analysis techniques, often referred to as text mining or computational text analysis (e.g., Ignatow and Mihalcea, 2018; Ignatow and Radev, 2017; Shu, 2020; Wiedemann, 2016). Particularly, the dissertation will position itself within the body of work that sees the potential of integrating quantified text data with thorough interpretations to develop sociological theory while at the same time stressing the limitations of both computational methods and digital data (e.g., Bonikowski and Nelson, 2022; Kang and Evans, 2020; Keuschnigg et al., 2018; Lindgren, 2020, 2019; Nelson, 2020, 2019; Spirling and Rodriguez, 2019). Worth mentioning is that this is a fast-growing new field for sociologists and that new research is constantly being published – as will be noted in the chapter on literature review. This study will not be the first of its kind, since the past few years have seen some studies dedicated to mining texts produced within the discipline for gaining empirical insights into the form and shape of sociological knowledge (e.g., Moksony et al., 2014; Moody et al., 2022; Vanderstraeten, 2010; Zougris, 2019). As will be shown in the literature review presented in the following chapter, there is, however, a tendency in these studies to accept commonsense understandings of sociology and not engage thoroughly with the theoretical insights on the discipline as a social phenomenon generated within the sociology of knowledge.

To perform computational text analyses capable of investigating the problem at hand, one would need empirical material that reflects the discipline’s understanding of what counts as proper sociology. While sociological journals and textbooks can be argued to fit this criterion to some extent – and have been approached in that very manner, as the chapter on literature review in this dissertation will show – it is debatable whether one can make the case that they are products that have survived the scrutiny of disciplinary representatives agreeing that ‘this text represents sociological knowledge’. In comparison, precisely such a disciplinary agreement is needed to issue a PhD dissertation in sociology – the cumulated product of the highest degree in sociology and the final stop in a journey that for many graduate students has taken decades. Dissertations are further relatively freed from the publication biases affecting, for instance, articles and contributions to anthologies, which have to fit the scope of their intended journal or book (cf. Fortunato et al., 2018). In the doctoral defense, as well as in seminars preluding it, repre-
sentatives of sociology (who are sometimes joined by faculty members in neighboring disciplines) critically assess whether the work presented lives up to the standards of a PhD in general, and is ‘worthy of the stamp of sociolo-
ygy’ in particular. In this sense, dissertations have a unique position in sociol-
ogy, with one foot in the scientific or research side of the discipline and one in the educational or teaching side. Dissertations in sociology have further been argued to be “charged with reproducing or transforming the field […] [and to] fuel the disciplinary lament of fragmentation” and will, thus, serve as this study’s empirical point of departure (Heiberger et al., 2021, p. 1167).

Countries have different legalizations for awarding PhDs and, in some cases, dissertations are never published and can therefore only be accessible to a handful of people. However, in a country like Sweden, all defended dissertations in sociology are printed by university presses or other publishers, and therefore accessible to the public’s eyes. This has been the case since sociology was institutionalized in Sweden in 1947 (cf. Larsson, 2001). Thus, it is argued that the corpora of dissertations in sociology, which in this case are defended at different universities in Sweden, provide ideal material for investigating what sociological knowledge is (not) (i.e., disciplinary ideas of the current state of sociology’s development as a science) with computational text analysis. Worth mentioning is that although the ‘local context’ in which these dissertations have been defended is without doubt interesting, and why this context is presented in Appendix A, this dissertation is not about Swedish PhD programs in sociology or the PhD degrees that they confer. Instead, this dissertation focuses on the question of what sociological knowledge is by deploying computational techniques on dissertations in sociology, which are hereby conceived as examples of knowledge that has been ‘approved’ by the discipline to be sociological.5

Aim

The main theme of this dissertation is to scrutinize the crisis of sociology by studying Swedish dissertations with computational tools for analyzing text. In the literature of the sociology of knowledge, this crisis is conceived as the fragmentation of sociological knowledge vis-à-vis the Kuhnian notion of paradigm, which has served as the central locus of how sociologists have theorized the state of their science. In this perspective, the problem of what

5 Although most studies studying sociology with computational methods rely on American or international data, relying on Swedish data is not deemed to be problematic for the task at hand. As Piotr Sztompka (2009) phrased it in One Sociology of Many?, even if it seems like sociology is becoming increasingly fragmented from a local point of view, globalization has made the discipline more unified than ever before when viewed with global eyes.
sociology is has been understood through the lens of what sociological knowledge is not. Thus, this dissertation aims to shed new light on the crisis of sociology by empirically scrutinizing prevailing disciplinary understandings of the state of its knowledge production through computational text analysis. To answer this aim, a series of computational text analyses resting on a clear theoretical foundation will be conducted on the corpora that constitute the empirical material for this study, which is dissertations in sociology.

Outline of the Dissertation

To investigate the aim stated in this chapter, the dissertation is henceforth comprised of nine chapters. Before diving into the specificities of each chapter, it might be helpful to clarify the general structure of the dissertation. Following this introductory chapter, there is a set of four chapters focusing on formulating the overarching research strategy, which culminates in Chapter 5 where the three research questions that will guide the empirical studies are presented. Thereafter, another set of four chapters is devoted to answering these research questions by presenting the results of the analyses conducted. The dissertation is closed off with a concluding discussion of the answers the dissertation can provide to its aim by relating the analyses conducted to the previous studies in the literature. In addition, it is worth making a note on the style in which the dissertation is written. The goal has been to write in a way that is understandable to the general audience of mainstream sociology and, particularly, comprehensible for the tradition referred to as the sociology of knowledge. With this follows an assumption that the readers have neither previous knowledge nor a deep interest in the computational procedures serving as the methodological backbone of the dissertation. Thus, to meet the potential needs of the more statistically inclined reader, the technical and mathematical aspects of the models used can be found in the appendices.

Following the increasing availability of digital sociological publications and developments in computational tools for processing them, Chapter 2 encompasses a literature review of recent trends in empirical studies within the sociology of sociology. The review focuses on what texts are called upon to represent sociological knowledge, how these textual representations are analyzed, and through what theoretical models the results are understood. The research questions presented in Chapter 5 are formulated to explore the potential evidence for these theoretical models by conducting computational analyses of dissertations in sociology. Drawing on insights generated from the literature review, Chapter 3 sets out to establish a theoretical framework
for studying sociological knowledge. The framework is focused on operationalizing Thomas Kuhn’s paradigm theory by combining its reception in the sociology of sociology, primarily the work of George Ritzer, with an early epistemological outline for tracing the existence of paradigms with computers presented by Margaret Masterman. Inspired by the research of Andrew Abbott on the discipline of sociology, it is suggested that the prevalence of paradigms might be investigated by locating repeated instances of a particular version of the proper research object of sociology. These versions are found in the form of poles in sociological dichotomies. Drawing on the early work of Piotr Sztompka, the chapter delineates nine of the most common dichotomies on the level of ontology, epistemology, and methodology to be sought when analyzing the dissertations.

Chapter 4 offers a methodological discussion of the potentials and limits of incorporating computational methods in sociological studies of knowledge. The path suggested entails overcoming ingrained methodological divisions in sociology by analyzing the outputs of the computational procedures through an interpretive perspective. The chapter also includes information on how patterns in the textual data are measured as well as how the dissertations in sociology defended in Sweden are sampled and how the two corpora analyzed are generated – one comprised of 380 dissertation full-texts (2000-2019) and the other of 815 abstracts for the dissertations (1980-2019). Again, the technical details of the computational techniques applied as well as the corpora are found in the appendices. Based on the argument running through the aim, the literature review, the theoretical framework, and the methodological approach, Chapter 5 formulates three research questions to be investigated in the corpora of dissertations in sociology. The first targets the fragmentation of sociological knowledge, the second the existence of paradigms in sociology, and the third the social conditioning of sociological knowledge. Further, an overview of how the research questions will be investigated by four overarching empirical studies, each constituting an analytical chapter in the dissertation, is given.

Chapter 6 and Chapter 7 make up the first half of the empirical part of the dissertation. Both chapters explore the three research questions on the corpus of abstracts for dissertations in sociology. Chapter 6 focuses on mapping out recurring ‘word constellations’ in the corpus, which come in the form of correlated and clustered word occurrences. Chapter 7 continues the investigation from the angle of ‘thematic arrangements’ generated with structural topic modeling, a technique for detecting latent topics and measuring the effect of metadata on their distributions. The main target of this part of the dissertation is to explore the research questions in light of the word constellations (Chapter 6) and the thematic arrangements (Chapter 7) generated in the analyses. That is, are these analytical constructs becoming more frag-
mented over time, reproducing certain sociological dichotomies (paradigms), and patterned across time and place?

The second half of the empirical part of the dissertation is constituted in Chapter 8 and Chapter 9. In these chapters, new angles on the three research questions are explored on the corpus of dissertation full-texts. Chapter 8 investigates linguistic similarities and differences of the most frequent words used in the dissertations to model linguistic styles. The styles are then scrutinized by looking into their most specific words and close readings of the most specific dissertations to investigate whether they resonate with certain positions in a set of sociological dichotomies. Chapter 9 is based on exploring to what extent the dissertations articulate divergent meanings of sociology, society (the main research object of sociology), and the sociological dichotomies presented in Chapter 3. The leading idea is that words can be assigned their meaning by the company of words they occur next to, which is explored by modeling word embeddings. In both chapters, the analyses also explore to what degree these representations of sociological knowledge are conditioned by time and place.

Drawing on the problems posed in the sociology of sociology literature, Chapter 10 is focused on responding to the aim by discussing the answers the conducted analyses can give to the posed research questions. This discussion involves addressing the empirical proofs that sustain prevailing understandings of sociology and suggest a potential way out of its supposed crisis based on the state of the art. Lastly, the suggested contributions of the dissertation are given and paths for future work in the sociology of sociology based on computational methods are delineated.
2 Reviewing Recent Trends in the Sociology of Sociology

As proposed in the introductory chapter, one of the most central issues for sociology is that the discipline is said to be in a crisis. The crisis of sociology seems to be rooted in a lack of a unanimous understanding of what sociology is and a supposed fragmentation of sociological knowledge. Yet, it remains unclear whether these prevailing understandings of sociology are reflected in the actual knowledge being generated in the discipline, i.e., sociological knowledge. Sociological knowledge is the object of study for the sociology of sociology – a branch of the broader tradition of the sociology of knowledge that emerged in the 1970s (cf. Merton, 1972). Pioneering works were primarily based on philosophical and historical studies of sociological knowledge, often focused on fathoming the reasons behind The Coming Crisis of Western Sociology (Gouldner, 1970) and the pathways for sociology to take if it wanted to avoid its self-destruction. With the increasing digitization of sociological publications and developments in tools for processing text as data, the last decades have witnessed the emergence of a ‘new wave’ of empirical research within the sociology of sociology. In line with the aim of this dissertation, this chapter explores these recent trends in the field, and the insights they bring, with a focus on textual representations of sociological knowledge. The review will explore how these studies conceptualize sociological knowledge, what type of texts they call upon for representing sociological knowledge, the kind of methods they utilize to study these textual representations, and the overarching theoretical models through which the production of sociological knowledge is conceptualized.

The guiding principle for tracking down the part of this ‘new wave’ in the sociology of sociology most relevant for the dissertation’s aim has been to set the inclusion criteria for this literature review to empirical studies of textual representations of sociological knowledge published between 1990 and 2019 written for an international audience (in English). To locate literature suitable for this inclusion criteria, Web of Science Core Collection (WOS) was consulted in January 2020. A broad topic search was then conducted – in WOS – to include studies mentioning either “sociology of sociology” or “sociological knowledge”. This resulted in a total of 166 items. All infor-
The information provided by WOS was then exported into a CSV\(^6\) file and explored with VOSviewer to evaluate the sample. This included computing networks constituted by words co-occurring in titles and keywords, as well as looking into citation trends. Via this method, it was easy to exclude papers that did not fit the inclusion criteria. Particularly, it was found that a good share of the papers were reflections of sociological knowledge rather than empirical studies.

The second step taken while conducting this literature review, which involves less than 100 papers, focused specifically on abstracts to determine whether or not they fit the inclusion criteria. Only about a third of the empirical studies included in this sample were found to target *textual representations of sociological knowledge*. To expand this sample through ‘snowballing’, the reference lists of the papers living up to the inclusion criteria were consulted in search for akin studies, which resulted in the addition of a handful of papers. Thirty-two peer-reviewed journal articles made it to the third step, where all papers were read in full-text and manually coded according to 54 variables.\(^7\) The variables were generated abductively by, on the one hand, consulting the aim of the dissertation and, on the other hand, inferring the main tendencies of the papers found when performing close readings. The same steps were once again performed in May 2023, resulting in 11 papers being added to the review, which were read and coded by the same variables, giving a total of 43 papers.

Before diving into the literature included in this review, it is worth mentioning the types of studies that are excluded from it. The focus on *empirical studies* entailed that all studies based solely on conceptual reasoning had to be excluded. Worth mentioning is that these types of studies make up a large share of the earlier sociology of sociology literature and include valuable insights into the condition of sociological knowledge, which is why some of the more influential theoretical works in the field will be addressed in the next chapter on theory. More contemporary examples include writings based on finding and solving theoretical puzzles such as, for instance, the intrinsic inconsistency in sociology derived from the works of Weber and Durkheim (Alexander, 1982), or how embracing interpretation can save sociology from

---

\(^6\) CSV denotes comma-separated values and is a delimited text file where each line of the file is a data record and each record consists of one or more fields, separated by commas.

\(^7\) The variables include the theory of sociology (main theoretical approach, referencing paradigm, referencing fragmentation/pluralism, and referencing methodological divide), the discipline(s) of study (solely sociology, including other social sciences, or including other sciences), their empirical material (the sampling strategy, the data types, and the number of units), the context of the study (national, international, and level of Anglo-Saxon focus), period of study (time span and year count), quantitative methods applied (descriptive, bibliometric, statistical, and NLP), and qualitative methods applied (close reading, manual coding, and historical method).
its troublesome inclination toward realism(s) (Reed, 2008). This also includes accounts of sociologists describing their own or others’ experiences of the discipline (e.g., Andersson, 2014; Ayres et al., 2009; Sokolov, 2019b; Twamley, 2015). Although interesting in and of themselves, they are not empirical studies as such.

Further, the exclusion criteria entail that many studies within the history of sociology and adjacent fields are left out. A common example is the literature on sociologists’ boundary work or other historical studies of the institutionalization of sociology as a discipline in the late 19th and early 20th century (e.g., Evans, 2009; Gaziano, 1996; Larsson, 2001; Meloni, 2014). Other examples include studies of sociology’s entanglements with the public sphere (e.g., Jøresten, 2008; Kropp, 2013; Kropp and Blok, 2011), biographical investigations of sociologists (e.g., Cunnigen, 2003; Heilbron, 2011; Wacquant, 2013), and the canonization processes of sociologists (e.g., Foy et al., 2018; Ginerskov, 2021; Outhwaite, 2009).

Further, the attention to textual representations of sociological knowledge means that contemporary studies akin to Social Knowledge in the Making (Camic et al., 2011), often focusing on the doings and sayings of sociologists, were also excluded. This includes observations of assessment processes and interviews with social scientists to unravel ‘how they think’ (Lamont, 2010) or what, according to them, can be characterized as ‘good sociology’ (Hokka, 2019), as well as the practices surrounding the production of sociological knowledge (Stacey, 2004). The exclusion criteria furthermore mean that a few survey-based studies were left out (e.g., Bowman et al., 2014).

Shifting the focus to the papers reviewed, the goal has been to get a broad impression of recent trends (i.e., from 1990) in the sociology of sociology, with the aim of fostering a general understanding of how these empirical studies represent, study, and conceptualize sociological knowledge. The analytical strategy has therefore entailed differentiating between the texts studied, how they are studied, and within what theoretical frame they are understood. Thus, this chapter is structured along four overarching sections. In the first section, which aims to answer the question of what characterizes the ways in which the sociology of sociology relies on texts to represent itself, the emphasis is on the actual texts that this literature deems to constitute sociological knowledge. This includes the type and number of texts, the timeframe of these studies, and the context of their samples. The second section asks what methods (quantitative as well as qualitative) these studies employ in order to study these textual representations of sociological knowledge. The third section aims to categorize how the studies conceptualize sociological knowledge by scrutinizing what kind of overarching theoretical models these studies actually apply. The results of each section are, last-
ly, discussed in relation to each other, with the aim of formulating a strategy for this dissertation, which is, as stated in the introduction, an empirical study of what constitutes sociology in a particular setting.

Textual Representations of Sociological Knowledge

This section of the chapter addresses the types of textual objects that are called upon in the literature to represent sociological knowledge. Based on this grouping, a few problems shared by the literature reviewed are found. These point out the advantages and disadvantages that are particular for each category of textual representations of sociological knowledge that the reviewed literature relies upon. These issues are discussed with the results of studies diverging from the overall trend. The main sub-questions that come into play in this first part of the review are the type and discipline of the publications in focus, the languages they are written in, and their national and multi-local contexts. The reason for including the latter is connected to the question of whether sociological knowledge is spatially conditioned, which will be explained in the third section dealing with how the papers conceptualize sociology.

Starting with the largest category, studies of sociology journals, sociological knowledge is approached as whatever research articles are published during a given timeframe in journals defined as “sociological”. In the sample used in this review, the number of units analyzed goes from only 30 articles (i.e., Teplitskiy, 2016) to 167,959 articles (i.e., Evans et al., 2016), with an average of a few thousand units. The temporal reach of these studies spans from three years (i.e., Varga, 2011) to 83 years (i.e., Vanderstraeten, 2010), with an average of almost 40 years. The principal period investigated is the decades around the turn of the 21st century. Hence, studies targeting sociology journal articles have, in general, a notable scope in terms of units and time, and are characterized by analyses of sociological knowledge in present times. The latter is undoubtedly partly an effect of the delimited period of the studies included in the sample (i.e., after the year 1999). Yet, the general contemporary focus also implies that the majority of the studies in this literature review do not focus on the historical dimension of knowledge production. Further, as will be explained in the following paragraphs, the geographical and linguistic contexts in which the articles are published are somewhat homogeneous.

---

8 Other texts like reviews or editorials are in almost all cases removed from the study sample.
By far, the most common empirical objects in this sample of literature based on sociology journals are articles published in the top Anglo-American sociology journals. These journals either serve as the sole focus of some studies (i.e., Ennis, 1992; Moksony et al., 2014; Moody, 2006; Moody et al., 2022, 2022; Schwemmer and Wieczorek, 2020; Teplitskiy, 2016) or are compared with equivalent journals in English-speaking countries (i.e., Collyer, 2014; Gartrell and Gartrell, 2002; Grothe-Hammer and Kohl, 2020; Platt, 2006; Seale, 2008; Zougiris, 2019). Alternatively, these Australian, British, or Canadian journals are the sole focus of the study (i.e., Collyer, 2013a; Crothers, 2011; Payne et al., 2004; Platt, 2007a; Watts et al., 2020). Thus, it seems safe to suggest that the lion’s share of the journal article studies reviewed analyze the Anglo-Saxon version of the discipline. This can partly be an effect of the sampling only including publications written in English. Yet, an increasing trend around the world is to publish in international journals, and there is evidence in this sample of studies with national journals stemming from other regions in focus, such as the Nordic Countries and the Netherlands (e.g., Aaltojärvi et al., 2008; Bjarnason and Síghusdóttir, 2002; Erola et al., 2015; Vanderstraeten, 2010).

Many of the other studies included in this sample gather their data from larger international search engines and databases such as Sociological Abstracts (i.e., Crothers, 2011; Moody, 2004; Moody and Light, 2006), Web of Knowledge/Science (i.e., Evans et al., 2016; Korom, 2019; Seale, 2008), and Social Science Citation Index (i.e., Korom, 2020; Schwemmer and Wieczorek, 2020; Varga, 2011). While the primary scope of all three search engines is articles written in English, where publications from the US and the UK make up an disproportionately large share of the databases (cf. Mongeon and Paul-Hus, 2016), few of the studies problematize the context of the articles. Thus, in this literature, the textual representation of sociological knowledge is empirically studied by searching for it in databases that by definition will bring attention solely to Anglo-Saxon sociology. Thus, the dominant trend for both studies based on the top Anglosphere journals, and

---

9 Particularly American Journal of Sociology, American Sociological Review, and Social Forces, which have for decades been referred to as the ‘big three’ or ‘core sociology journals’ (e.g., Gaston et al., 1979; Moksony et al., 2014; Oromaner, 2008).
10 These include the Australian Journal of Sociology, the British Journal of Sociology, and the Canadian Journal of Sociology. It is true that Canada is a bilingual country, but as is shown in Platt’s (2006, p. 208) study of Canadian sociology, “CJS [i.e., the Canadian Journal of Sociology] and CRSA [i.e., the Canadian Review of Sociology and Anthropology], in principle, publish articles in French, but in practice are overwhelmingly Anglophone”.
11 One exception is Gabriel Abend’s (2006) article in which he compares the top sociology journals in the US and Mexico.
12 The Nordic journals in focus are Acta Sociologica, Sosiologia, and Dansk Sociologi, whereas the ones from the Netherlands and Belgium are Tijdschrift voor Sociologie (Journal for Sociology), Sociologische Gids (Sociological Compass), and Mens & Maatschappij (People & Society).
the mentioned international databases, is to frame the article to be about sociology in general rather than that their results regard “only” Anglo-American or British sociology (i.e., Leahey and Moody, 2014; Moody and Light, 2006; Seale, 2008; Varga, 2011). This is one of the reasons behind this dissertation’s choice to focus on textual representations of sociological knowledge – that are the basis for granting a PhD in sociology in Sweden – as the units of empirical analysis (for insight into the national setting at stake here, see Appendix A).13

In the great part of the literature’s sample that focuses on journal article studies, exclusively sociology journals are analyzed, and, consequently, journals from all other disciplines are not within the scope of the study. Some may deem this unfortunate, presumably since it risks isolating sociological knowledge and overlooking its relationality. A counter-argument could be that a focus on the social sciences as such or leaning on an array of studies that could be deemed to be sociology – for instance, found in metadata retrieved from a digital database (a form of categorization that could in principle be based on a mix of authors, editors, archivists, and so on) – is to follow an unknown judgement about what counts as sociological knowledge. This line of reasoning has steered this dissertation to solely regard sociology dissertations, since the PhD in sociology (a product of multiple institutionalized forms of disciplinary scrutiny by internal and external peers), which is a textual representation of sociological knowledge, makes up one of the most clear-cut and objective ways to determine whether something counts as sociology or not.

Despite the general trend found in this literature review to solely include articles published in sociology journals, there are a few exceptions, including other sciences that have generated interesting hypotheses. For instance, Varga’s (2011) analysis of the semantic fragmentation in 5,852 ‘core sociological journal’ articles, 41,924 economics articles, and 33,416 biophysics articles published between 2006 and 2008 underline the value of disciplinary comparison. By contrasting the different disciplines, the results of this specific study suggest that natural sciences are less disintegrated than social sciences, where sociology holds the most fragmented position. A similar finding is presented by Evans and colleagues (2016) in their study of “paradigmaticness” founded on a corpus of 167,959 journal articles pertaining to

13 Worth noting, however, is that this is not a dissertation about Swedish PhD programs in sociology, or Swedish PhD dissertations in Swedish, per se. Rather, as has hopefully been made clear up to this point, the dissertations are utilized as data for investigating sociological knowledge and entering the international disciplinary dialogue on what sociology is. This is not to say that the national context does not matter, which is why Appendix A has been crafted to describe this context.
the social sciences – sociology,14 economics, political science, and psychology – and the natural sciences – biology, chemistry, physics, and mathematics – respectively. Based on their results, the authors propose that physics and chemistry are the most paradigmatic sciences, whereas sociology and political science are the least so.

There is, in other words, value in including neighboring social science disciplines when addressing what sociology is, based on textual representations of sociological knowledge, and to this effect it seems also appropriate to draw attention to Moody and Light’s (2006) comparative analysis of 1,657 social science journals sorted into disciplinary clusters. The results of their study support the idea of sociology’s interstitial position to more fixed disciplines such as economics and psychology. In addition, by looking at 107,943 articles listed in Sociological Abstracts 1970-1999, these authors can support the idea of the top American sociology journals’ centrality, as well as that sociology papers are becoming more clustered over time following the expansion of specialized sociology journals. Their results further suggest a second center for the British sociology journals, specifically Sociology, as well as the peripheral and isolated position of the Nordic journal Acta Sociologica. Further, an interrelatedness between sociology and economics in certain topics is suggested in Korom’s (2019) study of 537 wealth inequality research articles published from 1990 to 2017 where the articles form three major networks: one for sociology, one for economics, and the last shared by the two. Lastly, Seale’s (2008) comparative study of 4,745 articles published in sociology of health journals – including two social scientific health journals – and general sociology journals implies that the sampled generalist journals tend to be more theoretical and focus on the level of social systems (not individuals) in comparison to the specialist journals.15

Thus, besides sociology of sociology’s preferred empirical mode (i.e., the favoritism of dealing with a single discipline), the literature on sociology journals studies almost exclusively analyzes English language articles in top Anglo-American sociology journals without clearly addressing the limitations of the results (i.e., Korom, 2020; Schwemmer and Wieczorek, 2020). This means that these are the publications that form the main empirical basis for how sociological knowledge “in general” has been conceptualized in this literature. A clear exception to this pattern is studies dealing with a sub-field of sociology that often thoroughly discuss the limitations of their context, data, and analytical approach (i.e., Grothe-Hammer and Kohl, 2020; Watts et al., 2020). Similarly, studies that explicitly target national and language dif-

14 The sociology sample is made up of roughly 7,000 articles.
15 Five of the journals concerned are situated in the US, four in the UK, and one is international.
ferences are naturally keen on addressing contextual variances of sociological knowledge (i.e., Erola et al., 2015; Gartrell and Gartrell, 2002; Platt, 2006).

Most studies analyze publications from the Anglosphere, including the US, the UK, and/or Australia. However, the results of those focusing on comparing these locations suggest clear differences between these English-speaking countries. One example is Collyer’s (2014) comparison of 842 articles published from 1990 to 2011 in American, British, and Australian journals related to the sociology of health and medicine, which implies that all journals have a substantial national preference. Yet, American journals have twice as strong a tendency as the rest within Collyer’s sample to prefer citing intranational studies, only use local study materials, and not include a reflexive discussion of the study context. The latter suggests that the American studies in Collyer’s study tend to universalize their results.

A related national boundedness, here in terms of language preferences in cited references (i.e., we cite literature in the languages we speak), is found in Yitzhaki’s (1998) study of over 50 original research articles published 1985-1994 in nine leading sociology journals16 located in the US, the UK, Germany, and France. In addition, Platt’s (2006) comparative study of sociological research methods in 565 journal articles published in Canada, the US, and the UK managed to demonstrate some humble differences in national location and written language between English and French. Whereas the studies published in American journals and Canadian journals in the English language were more prone to be founded on quantitative methods, the articles found in the British journals and the Canadian journals written in French tended, to a greater extent, to use qualitative methods. As touched upon in the previous section, this opposition in knowledge production between the US and the UK is also found in other studies of journal articles in sociology, either on the level of methodology (Gartrell and Gartrell, 2002) or topical preferences (Zougris, 2019).

When looking at how knowledge is conceptualized in studies of international (English) sociology journals, Collyer’s (2014) thesis – even if she studies the sociology of health and medicine – seems to ring somewhat true for the sociology of sociology hereby in focus, since the studies do not explicitly engage in a thorough discussion on why their respective set of journals is fit to epitomize sociological knowledge. One explanation might be that the most

16 The American Sociological Review (US), Sociology (UK), The Sociological Review (UK), British Journal of Sociology (UK), Theory, Culture and Society (UK), Kölner Zeitschrift für Soziologie (Germany), Sozialwissenschaftliche Informationen (Germany), Revue française de Sociologie (France), and L’Année Sociologique (France).
prestigious journals in the discipline are presupposed to define sociology as a knowledge-producing discipline. Interestingly enough, this high trust in the top-ranked sociology journals is also manifest in studies of other objects. For instance, Cronin and Snyder’s (1997) study of the most cited works in the discipline uses book reviews published in journals for sampling (research) monographs in sociology. Consequently, sociology journals come to function as the principal objects for delineating what counts as a sociology monograph. Nevertheless, Cronin and Snyder’s (1997) results suggest that journals and monographs express two different worlds of citation patterns; a topic in and of itself that lies outside of the scope of this dissertation.

Another example of a related problem is found in Korom’s (2020, p. 1) study of “the prestige elite of sociology” in the 1970s and 2000s, which is based on a “newly generated text corpus of approximately 49,000 pages, which encompasses various genres of literature (encyclopedias, handbooks, journals, textbooks)”. At first sight, it seems as if Korom is trespassing the borders of sociology journal articles by including other literature. Yet, the book sample is here based on an index for monographs referenced in top sociology journals (cf. Cronin et al., 1997). Further, regarding the national bias, he takes the most cited authors in sociology from a primarily American data set of 15 sociology books and five sociology journals.17 Lastly, in his comparison of the reception of the prestige elites’ work in general, specialist and international sociology journals, Web of Science is the consulted search engine and American journals are prioritized.18 Still, Korom’s results suggest an interesting shift from a primarily American and educated upper-class elite toward a more European and socially heterogeneous elite that is recognized within general sociology, not only in a few specialties.

The importance of the publication type and the national setting for how knowledge becomes actualized is further articulated in the second category of textual objects used for representing sociological knowledge in the literature reviewed here, namely sociology books in the form of textbooks, monographs, or dissertations. The temporal focus of these studies is books published in sociology either around the middle of the 20th century or at the turn of the 21st century. Only one of the studies in this review’s sample considers another discipline, which is Korom’s (2018) comparative study of 30 text-

---

17 Two encyclopedias of sociology (one from the US, one from the Netherlands), two handbooks (one from the US, one from the UK), five top journals (three from the US, American Journal of Sociology, American Sociological Review, Social Forces, one from the UK, British Journal of Sociology, and one international, European Journal of Sociology) and ten text-books (nine from the US, one from the UK).

18 This analysis consists of 74,339 articles found on the Web of Science, which were published from 1956-2018 in 18 generalist and 36 specialist American sociology journals as well as 30 international sociology journals.
books in sociology, economics, and psychology, half of which were published in the 1970s and the other half in the 2010s. One striking difference between the studies of sociology books and studies of sociology journal articles or sociology departments is the wide range of units observed, ranging between four (i.e., Guth, 2008; Sand, 2008) and 5,540,000 (i.e., Chen and Yan, 2016), where about four out of five of the studies observe less than 80 units. One reason for a lower number of units might be how sociology books are produced. There are, for instance, many more sociology journal articles than textbooks, monographs, or dissertations in sociology. Perhaps more importantly, though, papers are much easier to access via digital databases (and have a higher chance of falling under an open-access agreement) than books.

Though not surprising, it seems important to point out that the majority of the papers that empirically study books in sociology as textual representations of sociological knowledge are based on studies that focus on textbooks. These studies partly follow the contextual bias of the journal articles since more than half focus on the Anglosphere, either solely (i.e., Harley, 2008; Keith and Ender, 2004; Korom, 2018; Platt, 2008) or partly (i.e., Schrecker, 2008), even if a significant number of studies deal with other countries, either in South America (i.e., Pereyra, 2008) or in Europe (i.e., Guth, 2008; Larsson, 2008; Sand, 2008). The results of some of these studies imply the interplay of national sociologies and more hegemonic forms of sociological knowledge, in particular, the influence of American sociology on more peripheral nations during the mid-century, for instance Sweden (i.e., Larsson, 2008), and Argentina/Mexico (Pereyra, 2008).

The general influence of American sociology is supported in a study of two strong sociology nations, France and Great Britain, by Schrecker (2008). By comparing the two countries a divergent pattern emerges, where the British sociology textbooks are far more empirical and focus on everyday life, while the French textbooks are in dialogue with philosophy and history, and build on a presentation style more common in literature. Schrecker (2008) also shows commonalities between French and British sociology, from which he speculates on the existence of a European sociology distinguishable from the American one. The value of looking for patterns of influence “below” the American is expressed in a study targeting a seemingly insignificant country in for sociological knowledge, such as Norway. Based on the study results, Sand (2008) makes an argument for the Norwegian influence on Nordic sociology through writing and translating sociology textbooks. The studies of sociological textbooks, thus, imply the value of studying contexts outside the Anglosphere to get a richer understanding not only of what counts as sociological knowledge but also how assumptions about this can stir our attention to what sociology ends up being within the sociology of sociology.
Something else worth mentioning is that by considering the national variety of the sample we find another explanation for why most of these studies only look at a few units. There seems to be a shared intention of these studies to explain aspects of sociological knowledge that go beyond what is expressed within the textbooks, which requires other types of empirical material to enter the analysis (sometimes included without commentary). About half of the studies only mention the sociology textbooks (i.e., Crothers, 2008; Harley, 2008; Keith and Ender, 2004; Platt, 2008; Schrecker, 2008), while the rest explicitly add curricula and reading lists for sociology courses or biographies of sociologists (i.e., Guth, 2008; Larsson, 2008; Pereyra, 2008; Sand, 2008). In all cases, however, there is an intention to describe the sociological knowledge of a whole nation. Further, while the results mentioned above suggest that sociology journal articles are becoming increasingly more clustered in terms of topics, the same does not seem to be the case for textbooks. A study of 35 American introductory sociology textbooks published in the 1940s and the 1990s by Keith and Ender (2004) suggests that the books share a similar structure but are highly dissimilar in terms of the concepts used (only a few percentages of the concepts were shared, and most were mentioned only once) and the way the sociology is framed as a discipline.

Lastly, there are two papers in the literature reviewed here that differ from those focusing on sociology textbooks, both in qualitative and quantitative terms. These papers share a focus on the Anglosphere and a large number of units, yet diverge in the type and sampling strategy of the books they investigate. First, the study conducted by Chen and Yan (2016) takes popular authors (e.g., “Erving Goffman”), sociological theories (e.g., “Social exchange theory”), sociology fields (e.g., “Urban sociology”), and sociological research topics (e.g., “Social capital”) – as its main type of entities. By looking into the English part of the Big Data Google Book Corpus (4.54 million books) between 1850 and 2008, the occurrences of these terms are compared and contrasted with each other. This is done, for instance, to show how often Parsons, social stratification, and historical sociology are referred to over time as well as how often such terms are used vis-à-vis the main terms of the other leading social science disciplines. Since Chen and Yan analyze massive non-sociological corpora, they are able to model the influence of these authors as signifiers of sociological knowledge in a multitude of different and general books to see how they are articulated over time. Yet, it is questionable how this general and equivocal corpus relates to the output of the discipline of sociology, as well as what the frequency of these terms in that corpus says about the development of sociological knowledge per se.

Second, we have a study conducted by Heiberger and colleagues (2021, pp. 1164, 1170) of 82,363 “sociology-related dissertations completed at U.S. universities” between 1980 and 2015, based on the criteria of whether or not
“the stem “sociol” was present in the title, abstract, keywords, or department”. The dissertations in the sample span over 12 disciplines and 275 university and college institutions, which, according to the authors, allows them to trace graduates from multiple disciplinary backgrounds who also contribute to the sociological field” (Heibeger et al., 2021, p. 1170). It is clear that this sampling strategy includes a very loose concept of sociology’s disciplinary boundaries since no apparent assessment is performed on the characteristics of either the departments or the disciplines in which the dissertations are defended. Further, the paper does not focus on sociological knowledge per se but sets out to “investigate how sociology students garner recognition from niche field audiences through specialization” (Heibeger et al., 2021, p. 1164).

Worth noting is that the only study of sociology books found and reviewed that specifically targets dissertations is that by Heibeger and colleagues (2021). While the paper suggests the potential of empirically scrutinizing dissertations for expanding our imagination of what sociological knowledge is, their own study, unfortunately, escapes this quality. The reason for this is that they neither limit themselves to PhDs in sociology nor take an interest in the content of sociological knowledge. This problem takes us to the last category of objects, where organizational structures are utilized to draw the discipline boundary of sociology.

Besides sociology journals and books, a small share of the studies reviewed focuses on departments of sociology as the site where sociological knowledge is produced, and specifically takes the research outputs of faculty members as their empirical object. One example that stands in line with the biases of Anglo-American articles addressed thus far is Leahey and Moody’s (2014) study of 1,785 journal articles published 1985-2004 that were written by the 181 tenure or tenure-track sociologists making up 99 PhD-awarding research universities in the US. Here, we are still stuck only with articles written by Americans; yet, sociological knowledge is shown to take part in a wider range of specialist journals not included in the obvious sociology journal canon. A similar framework is present in Bjarnason and Sifusdottir’s (2002) study that focuses on the citation patterns of the 1,205 articles published by the 271 faculty members of the 16 sociology departments in the Nordic countries,19 1981-2000. In this case, however, a non-American context is in focus and its results show that there exist both national and local publication patterns, suggesting that the single focus on Anglo-America might be problematic.

19 Denmark, Finland, Iceland, Norway, and Sweden.
Further, there are two other examples of focusing on the department level and prioritizing journal articles yet stressing the weight of examining different localities within a country. These are found in Collyer’s (2013a, 2013b) studies of 670 papers published by departments and journals constituting the sociology of health and medicine in Australian, which, according to the author, includes more general social scientific research outlets dedicated to health issues. The study suggests a number of “schools of sociology” in terms of theories and methods used at universities at certain times, and that the more prestige a university has, the less likely it is to have faculty employing sociological theories. Collyer concludes that the wealth and ranking of the university is the main factor determining the style of sociological knowledge.

The remaining papers focusing on sociology departments include journal articles as well as books, and present the argument that publication type matters. Two of these are a study of the aggregated publication practices\(^\text{20}\) of faculty members at 133 PhD-awarding sociology departments in the US by Moksony et al. (2014), and a study of the 1,038 publications present in Google Scholar by the 353 faculty of 2005 at the 16 Nordic sociology departments by Aaltojärvi et al. (2008).\(^\text{21}\) The results from Moksony et al. (2014) suggest that departments with a more quantitative “research style”\(^\text{22}\) tend to publish journal articles rather than books, while the opposite was true for departments with a more qualitative research style. The former departments performed the best in qualitative assessments, which the authors interpret as an indicator of a general preference for articles over books in sociology. The results of Aaltojärvi et al. (2008) also imply that the probability of publishing articles or books varies from department to department over time. Further, their analyses hint that different faculty positions have dissimilar preferences when it comes to publishing and getting different amounts of recognition (receiving citations and hits on Google Scholar) for their work.

Finally, there is one study with a broader definition of where sociological knowledge is produced that focuses on a context outside the Anglosphere and departments of sociology. De Haan (1997) considers the whole national output of sociology publications as expressed in the *Bibliography Dutch Sociology*, 1939-1952 and 1968-1987, which make up 4,893 publications.\(^\text{23}\)

---

\(^{20}\) Surprisingly, neither the number of publications nor the number of faculty members included in the study are presented in the article.

\(^{21}\) i.e., the same departments as in Bjarnason and Sifusdottir’s (2002) study presented previously in this chapter.

\(^{22}\) The indicator for research styles was taken from the 2007 *Guide to Graduate Departments of Sociology* and the websites of departments.

\(^{23}\) “This professional bibliography attempts to categorize as complete as possible literature (books, articles, and reports) on a yearly basis.” (De Haan, 1997, p. 198)
Although the study lumps together all publications into a single category, the results hint at a trend towards increasingly collaborative publishing. The analysis suggests collaborative networks within both Dutch sociology and international sociology, as well as collaborations with other social sciences such as demography, social psychology, and economics. Thus, in this part focusing on the sample of studies of sociology department publications, the overall trend seems to be that the publication type in focus matters, as well as the temporality of the national context and specific department where it is actualized.

In summary, this first part of the literature review, guided by the question of what textual representations of sociological knowledge undergo empirical investigations, has suggested the following: the studies predominantly analyze sociology journal articles published in the Anglosphere, especially Anglo-America. While these studies all offer valuable insights, the focus on a specific type of text in a particular context risks leading to a restricted conceptualization of how sociological knowledge becomes actualized over time and space. This is why it seems appropriate to offer the insights that were arrived at by analyzing the studies that compare publication types, nations, and disciplines.

First, there is purpose for including several local contexts in one study. One reason is that nations and departments tend to prefer different publications (cf. Aaltojärvi et al., 2008; Bjarnason and Sigfusdottir, 2002); for instance, departments doing more quantitative research tend to publish journal articles, while those dedicated to qualitative research have a leaning towards books (cf. Moksony et al., 2014). Second, previous studies have suggested a national boundedness of sociological knowledge that especially becomes problematic when the studies do not discuss the contextual limitations of their results. At least for Western countries, this includes a strong tendency to cite publications originating from the country of the journal (cf. Yitzhaki, 1998). The same seems to be true for the whole Anglosphere, but here the usage of local material tends to follow the same pattern, where American journal articles are the most national in this sense, as well as the least reflexive regarding their study context and the universal meaning of their results (cf. Collyer, 2014).

Thus, to expand the reach of the state-of-the-art empirical knowledge within the sociology of sociology, it seems important to include sociological knowledge produced in multiple localities outside the Anglosphere and articles published in the main sociology journals (housed in either the US or the UK). Particularly, there appears potentially to be value to be found by looking into books that tend to be overlooked in the literature, such as dissertations in sociology, and investigating how locality and history leave their
marks on the sociological knowledge produced. The fact that the previous studies of multiple departments have focused on the output of faculty leaves a door open for studying the knowledge products of the more precarious PhD students, who are a larger group of people and, presumably, bring a more versatile pool of research to the table. With a general idea of a proper empirical object for studying sociological knowledge in place, let us now move on to the methods employed in the reviewed literature.

Methods for Analyzing the Textual Representations

In this section, the focus point is on what methods are employed to analyze the textual representation of sociological knowledge in the literature reviewed here. This includes looking into how the methods might affect what types of results the papers present. The reason for dedicating a whole section to research techniques and variables stems from the lack of an empirical tradition in the sociology of sociology, which has been dominated by conceptual and historical studies. This prompts an outline of the methodological strategies pursued by the latest contributors to the field to circle promising routes for further studies.

A good share of the studies reviewed, 15 of the 43, do not consider actual textual representations of sociological knowledge but perform statistical techniques on variables connected to the objects in focus. These include the publications’ author, national context, and reference list. The most common methodological approach entails performing various types of bibliometric analyses where the entities are broken down into variables depicting the citation count the publications give away and receive. These can be used to explore patterns such as the language and journal preferences or co-author practices expressed in articles published in different countries (Vanderstraeten, 2010; Yitzhaki, 1998). Since this form of bibliometrics solely focuses on comparing co-authorship, citation scores, and cited references of scientific publications in sociology, it is often referred to as scientometrics (De Haan, 1997; Korom, 2019). In almost all cases, these patterns are explored further with one or two of the following three modeling techniques (in descending order of popularity in the sample): cluster analysis (e.g., Leahey and Moody, 2014), network analysis (e.g., De Haan, 1997), and multilevel modeling (e.g., Bjarnason and Sigfusdottir, 2002). The only exception to this trend in the variable-based studies is the paper by Aaltojärvi et al. (2008), which combines citation analysis with regression analysis.

More than half of the bibliometric studies only consider the identity of publications – including metadata such as the authors and their current depart-
ments or the name of the journal or publishing house – in relation to citation counts (Aaltojärvi et al., 2008; Bjarnason and Sigfusdottir, 2002; Korom, 2018; Moody, 2004; Varga, 2011; Yitzhaki, 1998). While such study designs offer valuable insights into the structures surrounding sociological publications, they do not focus on the content of the texts they study, but rather aspects ‘outside’ the texts, such as citation trends. Thus, the metadata or other types of external variables are in focus and therefore primarily the ‘material’ aspects of sociological knowledge. An exception is found in Crothers’ (2011) paper on research themes alongside citations and author affiliations drawn from British sociology journals. While focusing on thematical aspects of the journal articles, this study also solely relies on pre-existing variables, in this case, the 50-category classification scheme provided by Sociological Abstracts.

In contrast, a focus on content would entail looking into concrete textual data and analyzing the relations between the obvious patterns of the text which are, for instance, found in titles or keywords, and the less apparent formations of words expanding, so to speak, ‘between the lines’. There are some quantitative studies in the sample that not only include traditional bibliometric categories but also add various techniques for studying the actual texts deemed to represent sociological knowledge. One of these studies is a regression analysis of how the word length of sociology journal titles has changed over more than half a decade (i.e., Moody, 2006). A few studies take a step further into the texts to construct variables that are not assigned by standardized digital databases like the Web of Science. These include forming dictionaries to investigate how often different sociologists (i.e., Harley, 2008) and sociological concepts (i.e., Keith and Ender, 2004) occur in sociology textbooks (the latter generated from glossaries), the reception of sociologists in large corpora constituted by different types of sociology publications (i.e., Korom, 2020), and regression analyses of departmental preferences of publication types and methods (i.e., Moksony et al., 2014). In a seemingly more unconventional study, Chen and Yan (2016) construct an index through a principal component analysis of the occurrence of widespread sociology terms – names of prominent sociologists, sociological theories, sociology sub-disciplines, and research topics – found in a vast book corpus stretching over 160 years.

All papers addressed in the previous paragraphs can be said to empirically investigate terms derived from texts about the publications rather than the content of the publications. Presumably, it is hard to provide a substantial answer to the question of what sociology is by reading, as the saying goes, ‘what it says on the tin’. Rather, one would want to look inside the tin, and try to uncover the content of the texts. To elaborate further, categories derived from glossaries, databases, or constructed dictionaries can be seen as
sets of labels developed to express how texts ought to be read. While such labels are convenient for generating coherence (even in states of actual discord), their empirical bearing on the chosen textual representations is unclear and, consequently, so too is how sociological knowledge is actualized in the corpus.

In comparison to the more bibliometric or scientometric set of the literature review addressed above, about one-third of the studies have generated variables from qualitative analyses of the texts to explore trends based on descriptive statistics. For example, behind the results presented by Collyer (2014, 2013a, 2013b) and Platt (2006, 2007a), hundreds of journal articles were read and coded by hand. These qualitative investigations follow different coding schemes where the written language and explicit statements of whether a particular method, theory, or material is utilized, put a paper into one of a few pre-set categories. Also, Payne and colleagues (2004, pp. 156–157) hand-code the papers they investigate (with “a cross re-coding of a small unmarked sub-sample”) by research method, conceptual approach, and overarching topic (“using a slightly collapsed version of the British Sociology Association’s members’ interests list”), as well as by author characteristics (gender, seniority, and affiliation). This form of hand-coding is also conducted by Erola and colleagues (2015) on a smaller set of articles with several variables – theory, method, internationality, gender, education, and locality – that are utilized to perform nested multinomial logit models with methodology as the dependent variable. Similarly, Gartrell and Gartrell (2002) present a qualitative content analysis where each article is hand-coded according to an index constructed as the average of seven items representing positivism. This index is further investigated with descriptive statistics, ANOVA, and a rotated factor matrix.

Another version of this strategy is presented by Abend (2006) in a comparison of journal articles published in North and South America. Here the author constructs variables from reading the actual texts to denote the function of theoretical or empirical problems, ethical considerations, and statements of objectivity in the papers studied. An equivalent design is found in Tep-litskiy’s (2016) study of what aspects of 30 papers are consistent and contingent after peer review, which scrutinizes all papers in terms of text sections to show how, for instance, the number of variables used in the paper and the theoretical framing of the paper’s problem becomes modified during an assessment. It is true that all these hand-coding studies are based on finding ‘empirical regularities’ in the texts rather than utilizing pre-assigned labels to sort publications into categories. However, the patterns they map out appear to be derived from an arbitrary set of text excerpts (e.g., explicit references to theories, objectivity, and variables) that are, essentially, heavily at the mercy of what the researchers’ perceptions of what is deemed to be im-
portant prior to the analysis. In other words, unravelling the unforeseen and unanticipated aspects of these texts is not what was at stake when these studies were designed.

In addition to the studies mentioned up until now in this section, there are eight publications that, to an even greater extent, utilize a methodology commonly found within the humanities. The key method is a form of qualitative deep reading and text analysis that is applied for the sake of historical reasoning. The empirical material here comprises sociology textbooks (Crothers, 2008; Harley, 2008; Platt, 2008; Schrecker, 2008) that are read in tandem with curricula and reading lists for sociology courses and biographies of sociologists (Guth, 2008; Larsson, 2008; Pereyra, 2008; Sand, 2008). On this basis, the authors generate recurring themes that serve to establish a narrative of how local institutionalizations and developments, as well as international academic influences, have come to shape sociology within national or wider regional contexts. In a few of the cases, these qualitative text analyses involve simpler frequencies, such as references within and between the sampled textbooks (Pereyra, 2008), or the extent to which illustrations are used (Schrecker, 2008). Yet, the principal methodology in this sample is reading texts to inform a historical narrative.

The eight studies applying close reading and hand-coding put the actual texts under scrutiny at the forefront of the analysis and do not rely on labels provided by external systems, which is a good thing since these systems have, presumably, been constructed for other purposes (e.g., filing publications in a library based on general codes unfit for sociology as such). Further, the general pattern of these studies is to narrate overarching trends of how sociological knowledge evolves over several decades based on a small number of textual representations. This entails relying on historical factors ‘outside the texts’ (i.e., inferring knowledge external to the empirical sample) to form the narrative. Thus, one potential risk of such an approach is to reduce the width of the textual representations to only reach the limits of already established “truths” about sociology and its knowledge production.

There are, further, nine papers that also put the content of the texts supposed to represent sociological knowledge at the center yet analyze them using a computational methodology. This methodology entails employing techniques aimed at mapping out relations within a large corpus of texts, often spanning a concentrated time period (on average a thousand articles published over two decades). These studies are interpreted to be most closely related to the problem set out in this dissertation, as can be expected consid-
ering its aim (as stated in the introduction).\textsuperscript{24} Since some of these papers employ the type of computational methods that will be utilized in this dissertation, they will be described and discussed in the greatest detail thus far.

This selection of papers has been sorted into three sub-groups, based on the types of computational text analysis tools they apply, which will be explained one by one. In the first group, we find analyses of descriptive word frequencies (i.e., Korom, 2019; Seale, 2008; Watts et al., 2020). In one of these papers, Korom (2019, pp. 860–861) constructs clustered co-citation networks between a corpus of full-texts representing sociology and another, economics, to compare the corpora with “the simplest of all text-mining algorithms”, namely a “computer-aided content analysis” based on “purely descriptive word frequency counts”.\textsuperscript{25} These word frequencies are compared to explore the thematic distinctiveness between economics and sociology. Akin to this methodology, Seale (2008, p. 682) performs a “comparative keyword analysis”;\textsuperscript{26} based on the distance between words and the context of the publications (place of origin), on a corpus constituted by article abstracts published in seven top-tier general sociology journals and three top-tier sociology of health journals. Building on this comparison, a collocation analysis is performed to construct seven topics by “grouping together keywords into meaningful themes that are like the coding schemes of conventional qualitative thematic analysis”\textsuperscript{27} (Seale, 2008, p. 682).\textsuperscript{28}

The second sub-group of computational studies applies traditional statistical methods in innovative ways to analyze texts (i.e., Evans et al., 2016; Moody and Light, 2006; Schwemmer and Wieczorek, 2020). First, Evans and colleagues (2016) measure ‘paradigmaticness’ of journals in sociology compared to those of other social sciences and natural sciences, based on a corpus constituted by the titles and abstracts for each journal’s articles. This is performed by computing “the Shannon entropy of the language in each dis-

\textsuperscript{24} The aim of the dissertation is to shed new light on the crisis of sociology by empirically scrutinizing prevailing understandings of the state of its knowledge production with computational methods.

\textsuperscript{25} The analysis “ignores all extremely common terms … as well as all numbers” (Korom, 2019, p. 862).

\textsuperscript{26} In the study, the words are ranked using a log-likelihood test of the relative frequencies of words.

\textsuperscript{27} The topics are constituted by words concerning methods, concepts, place, divisions, health care, illness, and verbs/pronouns not adhering to one of the other six topics.

\textsuperscript{28} Echoing these approaches, the study of “Indigenous-focused journal articles” in two Canadian sociology journals, by Watts, Hooks, and McLaughlin (2020, p. 13), constructs themes from weighted word clouds of the most frequently occurring terms (in titles and abstracts) to compare how journals approach indigeneity over the course of several decades. The sample is made by querying whether titles and/or abstracts include words that ‘address Indigeneity’ namely: Indigenous, Indian, First Nation, Native, Aboriginal, Inuit, Eskimo, M\-\texttext{\texttext{\textemdash}tis}, and Savage.
disciplinary corpus” to measure consensus in word usage and the “cosine similarity to compare disciplinary discourse over time” (a relative measurement of overlapping words over time in two bodies of text), which is operationalized as rapid discovery at the periphery, and stability at the core of the discipline (Evans et al., 2016, p. 764. 767). Noteworthy is that the study is a comparison of word usage, so the words themselves are outside the scope of the paper. In contrast, Schwemmer and Wieczorek (2020, p. 9) apply “quantitative and automated methods of text analysis” to analyze abstracts from international sociology journals spanning across Europe and North America, which includes estimating key terms for each journal with $\chi^2$ independence tests. At the center of the paper lies the count-based “wordfish model”, a unidimensional scaling method based on “a Poisson distribution of word frequencies”, which is applied to illustrate a methodological divide in sociology by comparing the term weights and fixed effects of word occurrences (Schwemmer and Wieczorek, 2020, p. 10). Lastly, OLS regressions are applied to model the variability of time and publication outlets.

In the last article of the second sub-group, Moody and Light (2006) model a co-citation network of international social science journals to outline disciplinary boundaries, and another of only international sociology journals to map their intradisciplinary relational positions. The latter set of journals is further utilized to model count-based “paper topic networks” from the pre-processed titles, keywords, and abstracts adhering to the articles of four time-based sub-corpora²⁹ where “topic clusters” are generated based on the similarity of word vectors between papers (Moody and Light, 2006, pp. 68, 74). These are then visualized through ‘contour sociograms’ depicting the size and interrelation between (hand-labelled) topics to investigate the changing structure of the networks and the largest topics over time. Lastly, the authors test the modularity score of the four networks to illuminate the extent to which journal articles deal with the same topics. All in all, the authors interpret that a core of topics resides in sociology over time; a finding which this dissertation will put to the test.

From these two sub-groups applying computational tools to analyze text, it is clear that word frequencies can generate valuable insights into corpora such as overarching divisions, thematic differences, and consistency in language use. While the papers in the first sub-group primarily compare word occurrences across documents that are manually assessed, the second sub-group utilizes various statistical methods to investigate interactions between concepts by computing correlations between words and the effects of various metadata. One step further in the direction of utilizing computers to uncover latent patterns in the corpora, which might bring new insights into the form

and content of sociological knowledge, is found in the third sub-group, comprising papers that apply techniques from the fields of machine learning and natural language processing (i.e., Grothe-Hammer and Kohl, 2020; Heiberger et al., 2021; Zougris, 2019).

All three studies in this sub-group of computational text analysis generate latent topics and measure how they interact with different social and cultural characteristics of the publications (metadata) in sociology. In Zougris’ (2019) study of topical divides and bridges between American and British sociology journals, a dimensionality-reduction technique called latent semantic analysis is utilized to model 20 topics – labelled manually through their five most significant terms (varimax factor rotation) – on a corpus comprised of the titles and abstracts for the journal articles. Thereafter, a chi-square test of the distribution of topics referred to as “topic polarization index (TPI)” is developed to detect “polarizing and fusing topics” (Zougris, 2019, p. 71).

Further, a single correspondence analysis is utilized to chart the dependency of categorical variables to depict the distribution of the topics vis-à-vis their nationality (based on the TPI). The paper also includes a trend analysis of how the topics of ‘theory’ and ‘methods’ changed over time in both countries.

Grothe-Hammer and Kohl (2020, p. 423) employ similar content-analysis techniques to a corpus of international sociology journal articles in full-text to challenge the notion that organizational sociology is in decline. This involves hand-coding a sample of articles distributed over time and, through this sample, predicting the codes of the remaining corpus with a supervised machine learning technique called support vector machine. Through this methodology, a corpus of organizational sociology publications was generated on which an unsupervised machine learning technique called topic modeling was run to automatically generate topics based on the probability that words co-occur in documents. The study also features a citation analysis. A similar topic model is found in an analysis of specialization and career success in sociology by Heiberger and colleagues (Heiberger et al., 2021, p. 1174) that target “sociology-related’ theses” defended in the US. The authors fit a structural topic model of 60 topics on the abstracts of the dissertations and explore the probabilities that two topics co-occur in the same abstracts through network analysis. Structural topic modeling allows document covariates to improve the model and estimate the effect of independent variables on the topic-document distributions, in this case by constructing “metrics for students’ topic choices, focus, novelty, and consistency” (Heiberger et al., 2021, p. 1171). Further, the authors conduct an event-history analysis.

---

30 Sociology-related is defined as “all dissertations in which the stem ‘sociol’ was present in the title, abstract, keywords, or department” (Heiberger et al., 2021, p. 1170).
to predict “the likelihood of becoming an advisor, and time to first advisor-
ship” in relation to a selection of topics\textsuperscript{31} articulated by the abstracts, as well
as social characteristics,\textsuperscript{32} or the author and the metrics\textsuperscript{33} constructed for spe-
cialization and career success.

The review of the third sub-group suggests that machine learning can be
utilized to enable a fine-grained analysis of what the chosen textual represen-
tations of sociological knowledge are comprised of. This involves compari-
sions of the corpus and the document level, as well as how sociological
knowledge evolves over time and is distributed across space. Particularly, it
appears that it is fruitful to combine both unsupervised machine learning
techniques like topic modeling and latent semantic analysis, utilized for clus-
tering and uncovering latent topics, and supervised machine learning tech-
techniques like support vector machine or event-history analysis, enabling classi-
ification and regression.

This second part of the literature review, guided by the question of what
methods are utilized for analyzing textual representations of sociological
knowledge, has found three major categories in recent studies within the
sociology of sociology: statistical analyses of metadata and dictionaries,
hand-coding and narration-formation from close readings, and computational
text analyses. While all methodologies bring valuable insights into the condi-
tion of sociological knowledge, it seems that the computational text analysis
methods have the best potential to conduct both explorations and assess-
ments of prevailing understandings of sociological knowledge. This is be-
cause these techniques, drawing on both statistics and machine learning, are
the ones most prone to challenging the preconceptions of the researcher,
which is particularly important in a field like the sociology of sociology,
where sociologists study their own discipline. With this said, it has been
suggested that the combination of computational text analysis and traditional
methods, such as close reading, hand-coding, and investigations of metadata,
can bring about the deepest sociological understanding of texts (cf. Nelson,
2020). Worth noting, there is a tendency in the field for many computation
studies to rely solely on titles and abstracts instead of full-texts to represent
the empirical objects of study. However, the potential problem of this bias is
at this stage unclear.

\textsuperscript{31} i.e., identity, quantitative methods, race, and health.
\textsuperscript{32} i.e., gender and race.
\textsuperscript{33} i.e., focus, novelty, consistency, graduated from elite university, defended at a sociology
department, and publication count.
With this understanding of what texts are used for representing sociological knowledge and how they are studied, we can now move on to explore how sociological knowledge is conceptualized in the studies reviewed.

Theoretical Models of Sociological Knowledge

In this section, the focus lies on delineating the common grounds in the theoretical frameworks applied in the papers under review. To map out these frameworks, they are sorted into three overarching categories that represent different models for how sociology and its knowledge production are conceptualized in the literature. Before diving into this part of the review, it is worth noting the theoretical models do not form exclusive alternatives; quite the contrary, as it was found that many papers tend to combine two or three models in the same study.

In line with the introductory chapter, the lion’s share of the studies reviewed assumes sociology to be in a state of fragmentation or, to use a more positive term, pluralism. As will be shown in this section, fragmentation is imagined on different levels, from applied methods to epistemology and research topics. A modern cousin of the concept seems to have emerged in the term specialization, illustrating a process where sociologists are becoming increasingly focused on sub-fields and alienated from the disciplinary core. One of the most referenced sources for conceptualizing fragmentation is Abbott’s theory of sociological knowledge as changing through a fractal morphology, which will be described in relation to the studies applying it below.

The paper by Holmwood (2010, p. 650) provides one of the more historically informed readings of sociology’s fragmentation and “fragile state” from the 1940s, when Znaniecki stated that “sociology is disintegrating”, to the 21st century mode of “subspecialties drifting apart from each other”, making it seem as though “the discipline can only be pursued, if at all, against its own tendency for self-subversion”. In the reviewed literature, fragmentation is, in general, described as a ‘weak’ and ‘fractured’ discipline that is increasingly losing its ‘core’ and is constituted by separated, and sometimes rival, groups. As outlined by Moody and Light (2006, p. 68), one theoretical camp holds that “fragmentation brought about by cultural politics will cause the demise of the discipline”, while another argues that “the fragmentation of the discipline will lead to a more pluralistic science” that will be beneficial for its knowledge production. Collyer’s (2013b, p. 257) paper is an example of the latter when imagining sociology to be characterized by its “high regard for theory or methodological pluralism […] promoted as ‘core strengths’ and defining features of the discipline”.
In contrast, other studies assume fragmentation to be an issue, which can be seen implicitly in lines like “given the porous nature of content in sociology” (Moody et al., 2022, p. 30) or “the sociological discipline lacks a core knowledge structure” (Keith and Ender, 2004, p. 20). One more explicit example of the negative connotations of “the fragmentation of sociology” is given by Varga (2011, pp. 164–165, 173), for whom it is a “semantic fragmentation” characterized by a weak “consensus on concepts”, operationalized as a lack of shared references in a field that increases its “interpretive potential”.

On a theoretical level, the author presents (without empirically investigating) different causes of sociology’s fragmentation – the Kuhnian “multiparadigm” science with “incommensurable schools of thought”, the Abbottian lack of “commutative knowledge production” and fractal morphology of “fractal circles”, or the Turnerian thesis of “external social factors” that “distract sociology from cognitively and theoretically defined research problems” (Varga, 2011, pp. 168–169). These three major theories are recurring in multiple studies and will be explained in detail later in this review.

As proposed above, in the empirical studies conducted within the sociology of sociology, fragmentation is often associated, and sometimes even seemingly synonymous, with specialization in the literature reviewed. For example, in Korom’s (2020, pp. 2, 4) study of the prestige elite in the discipline, sociology is illustrated as a “pluralistic mosaic” and argued to be harder to depict than in other social sciences since it is “an internally balkanized discipline” with “many areas of specialization” and “a variety of literature genres”. Similarly, Leahey and Moody (2014, pp. 229, 247) state that their research “speak[s] to broader trends of specialization in science”, “fragmentation within sociology”, and “increasing specialization and fragmentation in sociology”. Heiberger and colleagues (2021, p. 1167) make the specialization of sociology the main target of their study since they argue “the specialization strategies of young sociologists” are “charged with reproducing or transforming the field” and, thus, to “fuel the disciplinary lament of fragmentation”.

In contrast, Vanderstraeten (2010) offers a reading that diverg-

34 Led by the hypothesis that economics and, even more so, biophysics are less fragmented, the results of Varga’s (2011) comparative study of the three disciplines suggest that sociological publications to the greatest extent neither share references in general nor cites outside their sub-fields (suggesting a lack of common disciplinary concepts).

35 Pluralistic mosaic is a term borrowed from Sztomka (2009).

36 In their study, specialization is positioned on the level of dissertation topics, which are conceptualized as adhering to different ‘Denkgemeinschaften’ (Fleck, 1935), a concept that will be addressed in the next section. Specialization is operationalized into “three antinomies” – where the first two address the topical choice of sociologists in their dissertations – tradition versus innovation, focus versus breadth – and the last if sociologists continue to work on the same topic in the post-doctoral phase – consistency versus change (Heiberger et al., 2021, p. 1166). The study results suggest that no specific set of topics are promoted in sociology but that the successful dissertations express “targeted specialization and novel combinations” –
es from the other by arguing that specialization is something of a necessity for scientific development and not a fragmentation per se. Rather, it is argued, the “rise of disciplines is connected with the formation of groups or networks of specialists” that “is connected with the emergence of ‘scientific communities’” (Vanderstraeten, 2010, p. 560). Specifically, these networks are said to be facilitated by the establishment of specialized journals.

Since fragmentation and specialization are processes, it appears logical that they ought to be conceived within the realm of time. While the papers seem to agree on sociology’s fragmented state, they seem to differ in how they imagine the past states of the discipline – whether it was once “unfragmented” or is only facing an upsurge in its previous state of disintegration. One of the more sophisticated processual theories of fragmentation references in the literature is provided by Abbott in his books Department and Discipline from 1999 and, in particular, Chaos of Disciplines from 2001. According to Abbott (2001), sociological knowledge takes the form of fractals, and similar oppositions recur through time on different levels (from the discipline as a whole to, for instance, a single department or a scientific sub-field).

One striking tendency is that the majority of papers referencing Abbott’s work neither discuss the theoretical underpinnings of the system nor seem to apply the theory in their study (i.e., Chen and Yan, 2016; Vanderstraeten, 2010; Zougris, 2019). Nevertheless, a thorough application of Abbott’s theory can be found in two studies by Moody (2004; 2006). In the first study, Moody (2004, p. 215) addresses “the lack of theoretical consensus in sociology”, in which “rapid growth also contributes to perceptions of fractionalization”. Leaning on Abbott, it is argued that “the nature of sociology creates permeable theoretical boundaries” – new ideas are constantly introduced on top of the old – and via a “‘fractal walk’ through the available idea space” the same theoretical frameworks are applied to “multiple empirical questions”, which entails that “specialization” does not necessarily generate

which at different times have meant writing about identity, race, and statistical modeling (Heiberger et al., 2021, p. 1185).

For example, Zougris (2019, pp. 64, 82) references the author to stress “the epistemological fragmentation of the field”, to speculate that it “can be informed by the epistemological variations”, and to conclude that the “fractal thematic landscape” is explained on the level of different national traditions. In a similar fashion, Vanderstraeten (2010, pp. 561, 564) refers to Abbott for stating that “scholarly journals” discern “the very constitution of scientific disciplines as they have developed in recent years” and that they “provide us with rich material for the sociological analysis of the ‘morphogenesis’ of sociology itself”. Further, Chen and Yan (2016, p. 873) borrow Abbott’s famous image of sociology as being “a caravansary on the Silk Road, filled with all sorts and types of people and beset … all of whom are bent on reducing the place to vassalage” only to address “the complexities of disciplinary advancement of sociology” and then leave it be.
divisions (Moody, 2004, pp. 216–217, 235). In a related study, Moody and Light (2006) reference Abbott’s work not only to emphasize that sociology is fragmented but as a direct argument against the Kuhnian notion of (normal) sciences developing through revolutions. Following their reading of the theory, sociology is depicted as a stream of ideas that “ebb and flow with few dramatic fissures” and where “victors resurgent ‘defeated’ sociologies”, which makes the discipline’s boundaries porous across time as well as vis-à-vis the other social sciences (2006, pp. 68–69). Based on this “idea structure of sociology”, Moody and Light (2006, p. 69) deduct that one “would anticipate shifts in dominant topics through time, with some residual effects persisting from previous eras”. By performing such a study of topics in sociology journals, they argue against a “general sense of disciplinary fragmentation” and for Abbott’s model of change through fractal cycles, yet hold that “the process [is] happening in a more rapid manner” (Moody and Light, 2006, p. 83). Worth noting is that while most topics changed over time, the authors suggest that a “core of work” (e.g., culture, organizations, social identity, social institutions, and social stratification) remains over the 40 years studied (Moody and Light, 2006, p. 82).

To recap, a tendency found in many of the studies reviewed is to assume the notion of fragmentation as a seemingly static aspect of sociological knowledge, without investigating how this process is articulated over time. Further, geographical variations of the phenomenon are not mentioned, making fragmentation appear as a global phenomenon. The studies adapting Abbott’s fractal theory, primarily Moody (2004) and Moody and Light (2006), were found to open up a fruitful path for studying the seemingly rapid fragmentation of sociological knowledge as a set of topics or research areas. Thus, this dissertation will take inspiration from these studies to examine the assumption of fragmentation by targeting multiple localities of a country not studied in the literature.

As was suggested in the introductory chapter, the impact Kuhn’s work *The Structure of Scientific Revolutions* (1962) has had on the disciplinary conversation in sociology can hardly be overstated. Thus, it comes as no surprise that Kuhn is commonly referenced in the studies of sociological knowledge reviewed. While paradigm theory is addressed in many papers in the literature reviewed, it is, however, seldom operationalized but mentioned only in passing, not seldom conflated with one or several concepts drawn from other theories (i.e., Ennis, 1992; Keith and Ender, 2004; Korom, 2018;

---

38 The results of Moody’s (2004, p. 235) study of collaboration networks partly reject this idea by suggesting that methodology (quantitative or not) and theoretical focus (positivist-empiricist or not) create a divide in sociological knowledge production. The ‘methodological divide’ implied by this result echoes studies addressed in the next section.
Leahey and Moody, 2014). A clear exception is found in the operationalization of ‘paradigmaticness’ offered by Evans and colleagues (2016, pp. 757, 759). Following Kuhn, they hold that a “paradigm is a guiding set of theories, methods, and questions” that provides a “coherent tradition of scientific research”, and a paradigmatic science is characterized by a “consensus over the core knowledge” and “rapid discoveries” in contrast to “debating the validity of old claims” (Evans et al., 2016, pp. 757, 759). Due to the status of the paradigm, the authors claim, there is a symbolic boundary that divides “the ‘hard’, high-status, paradigmatic natural sciences from the ‘soft’, low-status, preparadigmatic social sciences” with their “multiplicity of competing schools” (Evans et al., 2016, pp. 757, 759). However, according to Evans et al. (2016, p. 758), empirical studies that “measure and quantify paradigmaticness have gone poorly for the most part”. To address this gap, the authors lean on Collins’ interpretation of paradigm theory to operationalize ‘paradigmaticness’ as expressing three main characteristics: disciplinary consensus, stability at the disciplinary core, and rapid discovery at the discipline’s periphery.

The combination of paradigm theory and an epistemological division, such as ‘hard’ vs ‘soft’ science, found in the above mentioned study is a common idea in the reviewed literature. For instance, in the paper by Moksony and colleagues (2014, p. 18), they state that physics and chemistry have been regarded as the ‘hard’ sciences with the most uniform and standardized paradigm, whereas ‘soft’ social sciences like sociology have been described as “fragmented and debates about fundamentals are the rules rather than the exception”. Following Kuhn, the authors argue that a shared paradigm entails that “most ground-laying work can be spared”, making scientific communications compressed and fitting for journal articles, whereas the lack of a

---

39 There are several examples of under-theorized readings of the paradigm to be found in the literature. Summarized in one sentence, Ennis (1992, p. 259) reads Kuhn’s theory as merely the “duality between intellectual and social solidarity” and “shared models of scientific practice” between groups, but never mentions the paradigm concept per se. With this reading, Ennis launches a study of the connections among sociological specialties based on sociologists’ self-stated preferences. Similarly, Leahy and Moody (2014, p. 238) only mention the paradigm in passing, writing that “newcomers to a field are more likely to instigate paradigm shift”, a broad (and bold) statement seemingly without empirical support. Further, two studies of textbooks in sociology reference Kuhn in ways that appear to be contradictory: on the one hand, they contend that textbooks serve as “repositories of exemplars” from “the dominant paradigms of the day” (Korom, 2018, pp. 912, 919), and, on the other, textbooks supposedly present the discipline “not as one or more distinct paradigms but rather through the multiple perspectives of varied authors” (Keith and Ender, 2004, p. 20).

40 The rate of the ‘paradigmaticness’ index is explored in articles pertaining to four social science and four natural science disciplines, where sociology together with political science comes out as the least paradigmatic, and physics and chemistry the most. The results suggest that sociological knowledge is characterized by an absence of a cohesive conceptual language and, while not directly stated by the authors, that the discipline is indeed fragmented in comparison to ‘harder’ sciences.
paradigm entails that “everything always needs to be stated anew”, giving rise to long books and monographs (Moksony et al., 2014, p. 18). Drawing on the work suggesting that publishing format preference coincides with methodological style and/or departmental culture, they present the idea that each publishing format can be rated to specific “work styles” or “research styles (hard vs. soft, quantitative vs. qualitative)”, and fundamentally two diverging views on the “intellectual enterprise one believes sociology to be” (Moksony et al., 2014, pp. 1716, 1720).41

Indeed, a traditional figure of thought related to the paradigm that recurs in the literature is that of a divide between two competing forms of sociology (cf. Friedrichs, 1970; Lepenies, 1988; Mouzelis, 2000). While this divide comes in different versions, the most common reference is the methodological version, separating epistemologies drawn from the humanities and quantitative methods from those more closely related to the (natural) sciences and quantitative methods (i.e., Collyer, 2013a; Erola et al., 2015; Gartrell and Gartrell, 2002; Payne et al., 2004; Platt, 2006). This is best illustrated in the study by Schwemmer and Wieczorek (2020, p. 4), where sociology is conceptualized as “a low-consensus discipline” that is torn between two “rival camps”, “aligning themselves epistemologically either to the natural sciences or the humanities”, preferring either quantitative or qualitative research methods, and either “aligned to constructivism, logical induction and theory-building” or “to positivism, deduction and falsification”. In a single expression, this is referred to by the authors as ‘the methodological divide of sociology’, and, it is argued, causes a threat to consensus formation and successive formation in sociology.42 In contrast to this general trend, the notion of a methodological divide is criticized by Payne and colleagues (2004, p. 153), referring to it as “the dubious dichotomy between ‘quantitative’ and ‘qualita-

41 Based on their results, they show that departments of sociology with a tradition of using quantitative methods tend to publish in journals, and those favouring qualitative methods are inclined to write books. This connects to another version of this divide found in the work of Lepenies (1988) and picked up in the study by Aaltojärvi and colleagues (2008, p. 18), in which they suggest sociology to be “a hybrid of the scientific and literary traditions”, where each tradition’s “understanding of social phenomena” is prone to its own form of communication.

42 Worth noting is that the authors interchangeably refer to these camps as “epistemologically demarcated schools of thought” a la Fleck, and “dominant research paradigms” in the style of Kuhn, as well as “Paradigmatic alignments […] in] nationally embedded epistemic cultures” without discussing the implications of either theory (Schwemmer and Wieczorek, 2020, p. 4). Nevertheless, their focus is on the Mertonian “prestige economy” that is argued to produce durable hierarchies only if the science manages to install a “boundary demarcation”, which will be a question of the sociologists’ taste (Schwemmer and Wieczorek, 2020, pp. 5–6). The results of the study suggest that a methodological divide exists in sociology where each side has a preference for certain methods and topics, and for publishing in their “own” journals that, according to the authors, is “mirroring the entrenchment between different paradigms” (Schwemmer and Wieczorek, 2020, p. 16).
tive’ methods’ that might be suitable for short commentary but ‘ultimately unsustainable and a poor basis for serious debate’. 43

To expand on these tendencies, it seems that the paradigm model still lingers on in the sociology of sociology by affecting how sociological knowledge is theoretically conceived and empirically studied. In general, the papers addressing paradigm theory seem to agree that sociology is not ruled by a single paradigm. In contrast, the most common reading appears to be the divide between two archetypical versions of sociology: one founded on the humanist or qualitative tradition and the other on the scientific or quantitative tradition. The notion of such a divide is not new but is suggested in Lepenies’ (1988) classical historical study to lay at the very foundation of sociology. This methodological divide is not without its share of critique (e.g., Payne et al., 2004) and it varies whether none, only the scientific tradition, or both are seen as proper paradigmatic candidates. This connects to the most pressing issue exposed in the review of theoretical models applied in the study, namely that the paradigm concept needs to be properly operationalized. Here the paper by Evans and colleagues (2016) makes a significant contribution. Now that their study has concluded that sociology lacks ‘paradigmaticness’ vis-à-vis other disciplines, the next step seems to be investigating the content and the form of sociological knowledge, which is outside the range of their study. Thus, this ‘research gap’ in relation to how paradigm theory is utilized in the literature will be one thing addressed by this dissertation. With such a theoretical framework and methodology, it would be possible to investigate whether there is evidence for the common assumption of a dichotomous divide in sociology.

In his classical book, Ideologie und Utopie, published in 1929, Mannheim (1960, p. 237) stated that “the sociology of knowledge has set itself the task of solving the problem of the social conditioning of knowledge” and concerns itself with “the significance of the non-theoretical conditioning factors in knowledge”. This seemingly simple but powerful idea, that knowledge emerges in relation to its societal surroundings, has been a key notion in the histography of sociology and is a recurring theme in the studies reviewed here. A common assumption in many papers is that sociological knowledge emerges vis-à-vis the context of its production, both in terms of history and geography. In these studies, the temporality of sociological knowledge is operationalized in a rather straightforward manner as depicting historical

43 Nevertheless, the results of their study of British sociology journals suggest a strong preference for publishing articles applying a qualitative rather than a quantitative methodology. According to the authors, sociologists tend to support the idea of “methodological pluralism”, but their analysis shows clear “anti-quantitative and anti-positivist methodological positions”, which suggests that sociologists are more tolerant in their opinions than their practices (Payne et al., 2004, pp. 154–155).
trends in, or the evolution of, knowledge. Conversely, spatiality is conceptualized in diverse ways, spanning from the global level, assuming differences in knowledge produced in, for instance, the global south versus the global north, through regional and national versions of knowledge, to the local level such as a specific place or site. In reviewing the social conditioning of knowledge in the literature, the strategy will be to move from the widest to the narrowest versions in the literature. This entails beginning with global and international conditioning, continuing with transnational and national conditioning, and closing off with local conditioning.

The papers addressing the global level appear to share the worldview that globalization not only reflects the current time but also affects sociological knowledge in various ways. A recurring theory is that of institutional isomorphism, particularly the one developed in the work of Heilbron (cf. 2014), stating that internationalism is increasingly making sociology around the globe more alike by sharing the same intellectual references, which is causing the ‘nationally specific’ traditions of sociology to fade out. The paper by Zougris (2019, p. 65) is based on a version of this theory, focusing on how the globalization of sociology could foster a “structural thematic isomorphism”, which is assumed to require a “high level of consensus in the core theoretical and methodological paradigms in national and international sociologies”. Zougris’ study of topics in American and British sociology journals suggests two “dominant national sociological traditions” that nonetheless to some extent share a common topical ground (Zougris, 2019, p. 82).

A more ontologically burdened assumption of the world found in the literature is the world-systems theory (cf. Wallerstein, 2004), holding that the global world is structured by an intellectual center or core and one or a set of peripheries, and that knowledge is conditioned differently based on where it is produced along this core-periphery axis. For instance, Collyer’s (2014, p. 253, 2013b) studies are based on “a union of the sociology of knowledge and theories of globalization” from which the author proposes a “world system of knowledge production and exchange” focusing on a dominant core and dominated peripheries in sociology. Sociological knowledge is here conceived to be situated in local institutions that are conditioned by much broader social contexts of power stretching to the global level. Collyer (2014, pp. 254, 264) argues that knowledge production in sociology is constituted by a self-referencing “core”, represented by the Global North in general and the Anglo-American and British countries in particular, that mainly builds on local materials, and a “periphery”, which reference results from the core and disregards local materials, such as African countries of the Global South. This leads to a “global, knowledge-based form of capitalism” with “developing countries being net importers of knowledge” (Collyer, 2013b, p. 246). An intermittent position, and, according to Collyer (2014, p. 253), a
much less studied one, is found in countries of the “semi-periphery” (the author mentions Russia and Australia), which are “open to exploitation by core countries while simultaneously able to trade to advantage among the countries of the periphery”. This dissertation utilizes the case of Sweden, which within this global system framework could also be conceptualized as a semi-peripheral country.

Around the turn of the 21st century, a common figure of thought has been that sociology suffers from methodological nationalism by structuring its studies within the frame of the nation-state, when conducting both small case studies and large comparative studies, despite the fact that globalization is said to entail an increasing rupture of national borders. Indeed, despite the global perspectives on the conditioning of sociological knowledge presented in the literature, studies continue to emphasize its national conditioning in present times (e.g., Crothers, 2011). For instance, in the paper by Zougris (2019, p. 64) addressed above, it is acknowledged that social conditioning takes different forms, and that “the geography of sociological knowledge” reaches from national traditions through organizational affiliations to research networks. The study results, comparing US and UK sociology journals, suggest that the two differ in topic preference, where the former prioritizes methodology and the latter theory, but also manage to show that they share a few topics such as Gender, Family, and Politics. This division within Anglo-Saxon sociology is a common figure of thought that re-emerges in several other studies under review (i.e., Grothe-Hammer and Kohl, 2020; Platt, 2006; Seale, 2008).44 Most interestingly for studies of sociological knowledge, the results of these studies suggest an increasing unity in conceptual frameworks applied within the sub-field, which speak against the common idea of fragmentation of theoretical approaches within the discipline (cf. Turner, 2010).

Possibly the most painstaking attention to the idea of national traditions is found in the paper by Abend (2006), which from the start presumes that so-

44 Platt’s (2006) comparative study of sociological research methods in Canada, the US, and the UK demonstrates some humble differences in national location and written language (English and French). Whereas the studies published in American journals and in Canadian journals in the English language appeared more prone to be grounded on quantitative methods, the articles found in British journals and those written in French in Canadian journals tended to apply qualitative methods. Another version of the US/UK division is found in studies extracting testable hypotheses from theoretical debates and historical typologies found in the genesis of a particular sub-field. One example is Seale’s (2008, p. 692) comparison of the intersecting histories of the sociology of medicine and sociology of health in the UK and the US respectively, in which they find “marked divergences between the sociological traditions of each country”. Another is the paper by Grothe-Hammer and Kohl (2020), a study of organizational sociology that exposes different national trends, where articles published in US journals generally engage in the sub-field and those in UK journals do not.
ciological knowledge production in the US and Mexico is governed by diverging epistemological beliefs. According to the author, the countries ought to be conceived as two “Denkgemeinschaften” (see Fleck, 1935) since they are governed by contrasting intellectual histories, are embedded in dissimilar cultural and societal institutions, communicate in different languages, and have diverging “thematic, theoretical, and methodological preferences” (Abend, 2006, pp. 2, 30). The results of the analysis suggest that US sociology, to a greater extent, uses middle-range theory, more deductive reasoning, poses propositions that are close to nomothetic, applies objective and passive language with mathematical symbols, and calls for ethical neutrality. In contrast, Mexican sociology tends to couple theory and data into joint reasoning, to use a more ideographic style of reasoning where the subjectivity of the writer is displayed, and to expose value judgments. On this basis, the paper concludes that the countries represent two “Styles of Sociological Thought” that cannot be translated and are incommensurable with one another (Abend, 2006, p. 29).

The design and results of the studies addressing the national conditioning of knowledge in the US and its “others” can be related to the narrative of positivism that is commonly found in Western sociology. This maintains that the discipline was once deeply influenced by the positivist movement following the end of World War II, which is supported in historical studies spanning from large and central countries like the US (i.e., Turner and Turner, 1990) to small and peripheral countries like Sweden (i.e., Larsson, 2001). Leaning on evidence that “theory circles in sociology had declared positivism to be dead” by the year 1990 and the assumption that “quantitative methodology is itself a defining characteristic of instrumental positivism”, Gattrell and Gattrell (2002, pp. 639, 645) specifically analyze the prevalence of positivism in the US and the UK. The study suggests that while positivism and quantitative methodology have started to vanish in the UK, they still linger on in the US, suggesting the emergence of two diverging national traditions of sociology.

If bracketing the fact that Anglo-Saxon countries serve as the sole empirical basis here (with the exception of Abend’s study of Mexico), which calls for analyses of other countries like Sweden, these papers seem to support the national conditioning of sociological knowledge. However, they do not address the problem of globalization, holding that the US and the UK in turn only reflect the dominant center of international sociology. In the literature, the dilemma between a global and a national understanding of sociology is probably most thoroughly theorized by Bjarnason and Sigfusdottir (2002) when scrutinizing the phenomenon of ‘Nordic sociology’. In their paper, they follow Wallerstein when postulating that “sociology is fundamentally a global enterprise” with a slow and uneven internationalization and
acknowledge the existence of “national sociologies” by stating that sociology’s growth has been “shaped by national ideals, social and political values, and accepted patterns of social organization” (Bjarnason and Sigfusdottir, 2002, pp. 254–255). By studying the citation patterns on the local level, in their case the output of 16 departments, they argue that each department has “a distinct intellectual identity” that contributes to the overarching constitution of national and transnational communities of sociology alike (Bjarnason and Sigfusdottir, 2002, p. 255).45

Regarding the local level, the paper by Collyer (2013a, p. 339) on sociological knowledge production at Australian universities offers a complementary angle by introducing “schools of sociology”, where a school signifies the institutionalization of a “specific orientation to practice” or a “shared theoretical or methodological approach to knowledge production”. The concept stems from a reading of the sociology of knowledge as properly addressing “the social determination of formal knowledge” as an entity produced by agents embedded in institutions that in turn are “influenced by social structures, cultures, and everyday practices” (Collyer, 2013a, p. 338). The study results suggest that schools of sociology are “institutionally and geographically-tied” and structured around methodology, particularly so for quantitative sociology since it – in contrast to the small and loose “theoretical networks” of qualitative sociology – follows “the natural science model of team work” and requires “substantial, on-going, in-house expertise” and, thus, to ensure the passing of “inter-generational expertise” (Collyer, 2013a, p. 349).46

When summarizing the adaptation of the social conditioning of sociological knowledge, it appears that a comparative study on the local level might well carry the potential to provide new insights into its production without presuming characteristics based on national histories or positions in a global

45 Based on their study of departments of sociology in the Nordic countries, Bjarnason and Sigfusdottir (2002, p. 267) found that half of their publications were in general sociology outlets (i.e., mainstream international sociology journals) and the other half in specialist journals. The authors speculate that individual trajectories are bound up in institutional ones where “the collective efforts of faculty may lead certain departments to become known as ‘powerhouses’ in the discipline” (Bjarnason and Sigfusdottir, 2002, p. 265). Worth mentioning is that this particular thesis is partly falsified by a study of citation patterns based on the exact same departments conducted by Aaltojärvi and colleagues (2008, p. 20), suggesting there “do not seem to be any Nordic sociology centres of excellence that attract recognition to their faculty”. Unfortunately, neither study offers an analysis of what department goes with what fields of specialization and, thus, it is unclear whether the claim that departments have a distinct intellectual identity is supported or ought to be rejected.

46 While Collyer’s (2013) study looks into the evolution of theoretical frameworks, no effort is made to scrutinize, for instance, what theory or theme goes with what place or locality, which its focus on schools of sociology as geographically and institutionally embedded would imply.
system. This is why this thread will be picked up in this dissertation, focusing on departments of sociology located in Sweden to explore the social conditioning of sociological knowledge by studying how it is produced at multiple localities within a single nation over time.

When wrapping up this section, it is proposed that the studies reviewed rely on three major models of sociological knowledge: fragmentation (including fractalization and specialization), paradigms (including scientific divides), and the social conditioning of knowledge (on a global, national, and/or local level). Before moving on to the concluding section of this literature review chapter, it seems appropriate to once again stress that not all papers reviewed here have been informed by a single theoretical framework. Some deploy two or more models for conceptualizing sociological knowledge. One telling example is Zougris’ (2019) study of topic preferences in different countries, which addresses fragmentation and fluidity (here by referencing Abbott), various versions of social conditioning (drawing on Mannheim, Kuhn, and others), as well as the institutional isomorphism in sociology (that national sociologies a becoming more alike). This study argues that the results support all three models. This is noteworthy since one might perhaps expect fragmentation to entail a topical increase, institutional isomorphism a topical unification across nations (presumably a decrease in the number of topics), and social conditioning that certain combinations of topics are nationally specific (in direct contrast to institutional isomorphism). This example illustrates a tendency found in the literature that sociological knowledge is not systematically theorized. Rather, much like the critique of sociology by which many studies are founded, concepts are used in a more sensitizing way. This suggests that the field could gain from a more systematic take on theory when studying sociological knowledge.

Nevertheless, the models of sociology outlined in this section show that a good amount of thought work has already been put into formulating and exploring some of the prevailing understandings of sociology by conducting empirical studies of texts. In this sense, the papers reviewed form a solid foundation upon which this dissertation can build further. Given that all models seem to not only still be presented within the sociology of sociology but also have their individual merits for our current understanding of sociological knowledge, it is proposed that this dissertation will have to address all three to engage with these recent trends in the field. Thus, discussions of how each model will be investigated will be needed. These are given in the next section, which summarizes the patterns proposed to be prevalent in the studies reviewed, and proposes the path forward for this dissertation.
Towards a Computational Study of Dissertations

This literature review of recent trends in the sociology of sociology, focusing exclusively on empirical studies of textual representations of sociological knowledge, has pinpointed several tendencies in the studies under appraisal. Based on the first section of this review – which brought attention to the kind of textual representations of sociological knowledge that empirical studies within the sociology of sociology utilize, it was found that to capture the complexity of sociological knowledge, a study ought to address a sizeable corpus (>100) that can account for both temporal trends and spatial variations. It was further suggested that there is a bias in the literature towards studies of sociology journal articles published in the Anglosphere (the US and the UK). Building on the results of the studies dealing with other texts (cf. Heibeger et al., 2021; Keith and Ender, 2004; Korom, 2018) and contexts (cf. Abend, 2006; Collyer, 2014; Erola et al., 2015; Pereyra, 2008; Yitzhaki, 1998), this bias might lead to a restricted understanding of sociological knowledge. Particularly, books and monographs tend to be overlooked in the literature, as well as countries between the core and periphery of the global world system. Further, the review highlighted that insights might be gained by addressing several local contexts of nations in a single study (cf. Aaltojärvi et al., 2008; Bjarnason and Sigfusdottir, 2002; Collyer, 2013a; Moksony et al., 2014).

Thus, to expand the reach of the state-of-the-art, it is hereby proposed that a study of the hundreds of dissertations in sociology defended between 1980 and 2019 at the five main departments of sociology in Sweden might bear the potential to add new insights into the production of sociological knowledge.

In the second section of this chapter, the focus was on the methods utilized for studying these textual representations of knowledge. It was there suggested that recent empirical studies within the sociology of sociology have primarily applied three categories of methods: statistical analyses of metadata and dictionaries, hand-coding and narration-formation from close readings, and computational text analyses. In addition, that section suggested that all methods have their particular merits, and might even bring about synergic effects when combined, yet that the best candidate for bringing new insights to prevailing understandings of sociology can be found in employing computational text analysis techniques (i.e., Evans et al., 2016; Grothe-Hammer and Kohl, 2020; Heibeger et al., 2021; Korom, 2019; Moody and Light, 2006; Schwemmer and Wieczorek, 2020; Seale, 2008; Watts et al., 2020; Zougiris, 2019). This is because these methods process the actual texts of the object chosen to represent sociological knowledge (instead of only metadata)
and can expose latent patterns in large corpus that are hard to see with human eyes. Thus, they have the potential to counter the prevailing understandings held by the sociologist of sociology. With this said, it was suggested that a combination of computational text analysis and traditional methods might be able to provide the best ‘sociological’ understanding of the texts studied (cf. Nelson, 2020), which explains why this dissertation’s aim is formulated as it is.

Further, it was found that the studies under review tended to utilize abstracts instead of full-texts as representations for the sociological publications under scrutiny. However, the potential implication of this bias is, to my knowledge, unknown. This suggests that there might be potential in analyzing full-text to expand our understanding of sociological knowledge that ought to be explored further. To be able to enter into a fruitful dialogue with the results of studies targeting abstracts, and the valuable insights they have proven to bring about, this review has suggested that this material ought to be included besides abstracts, and so both abstracts and full-texts ought to be explored in future empirical endeavors in the sociology of sociology. Thus, to live up to this condition, the study design presented in this dissertation will include one corpus of abstracts and another corpus of full-text versions of dissertations in sociology.

The third section of this literature review aimed to delineate the main theoretical models for conceptualizing sociological knowledge utilized by the studies. This led to the proposition that the literature relies on three major models of sociological knowledge: fragmentation (including fractalization and specialization), paradigms (including scientific divides), and the social conditioning of knowledge (on a global, national, and/or local level). It was further suggested that the three models reflect different trends in the sociology of sociology that all contribute to the current understanding of sociological knowledge, which is why it seems appropriate to design a study that is informed by, and operationalizes, the three theoretical models. This is why this dissertation’s research questions have yet to be formulated. Suffice to say that they have been crafted to respond to these three theoretical models, which is why they are formulated after the theory informing this dissertation has been spelled out, and after the methodology employed is presented. Thus, it is in Chapter 5 that this dissertation’s research questions are first formulated.

Given the proposed centrality of the models in recent trends in the sociology of sociology, this chapter will end with discussing how each theoretical model is understood and will be approached in this dissertation. While most sociologists would probably agree, at least to some degree, that their discipline is pluralistic, and probably even in a state of fragmentation, it is most
often uncertain what is concretely meant by this assumption. Therefore, it seems crucial to describe how fragmentation is understood in the field before conducting a study of sociological knowledge. Following this literature review, fragmentation often appears to be synonymous with the specialization of research topics, manifested in phenomena like an increasing number of sub-fields in sociology. Two of the more theoretically refined papers, i.e., Moody (2004) and Moody and Light (2006), lean on Abbott’s (2001) fractal theory to understand the seemingly organic way research topics in sociology can appear, disappear, and later reappear. The latter study of abstracts in an international social science journal suggests that most topics in sociology are gradually exchanged but that a core of work remains over time (Moody and Light, 2006). Thus, this fragmented development of sociological knowledge through thematic replacement ought to be investigated in other contexts, as will be carried out in this dissertation.

From the review, the fractal model appears to join forces with the paradigm model by rejecting the traditional ideal of science as ‘the linear progress toward objective truth’, yet diverges from it by conceiving the evolution of sociological knowledge as simultaneous processes of re-mapping instead of through sudden ruptures and revolutions. Yet, in the same manner, the existence or non-existence of one, two, or multiple paradigms in sociology cannot be taken for granted and needs to be empirically scrutinized. In the literature, the most prevalent version of the epic battle between paradigms in sociology is undoubtedly that of an alleged divide between soft knowledge or qualitative methods and hard knowledge and quantitative methods. While several studies seem to sustain such a divide, the relationship between the actuality of two methodological camps, tribes, or tendencies and the actuality of paradigms in sociology remains immensely uncertain. One could easily imagine a ‘normal science’ working under a unified paradigm with techniques stemming from vastly different methodologies. Here the analytical reception of the paradigm concept presented by Evans and colleagues (2016) can serve as a point of inspiration. Importantly, their study also manages to empirically sustain common assumptions such as that sociology is less ‘paradigmatic’ than economics and much less so than chemistry and physics. However, the question of how this lack of ‘paradigmaticness’ is articulated in sociological knowledge lies outside the scope of their paper and would presumably require a complementary operationalization of paradigm theory. Thus, this dissertation seeks to respond to this call by formulating such a theory in the next chapter (Chapter 3).

Lastly, regarding the social condition thesis, this literature review has proposed that the context of sociological knowledge is of great importance. Yet, rather than deductively postulate the effect of national histories or positions in a global system, it suggests that it might be approached inductively by
targeting multiple localities, which in turn form the building blocks of the national and global levels. Particularly Collyer’s (2013a) study opens up a path for investigating whether and how ‘schools of sociology’ might be institutionalized within departments of sociology that, building on Bjarnason and Sigfusdottir (2002), can set the tone for each respective intellectual identity and collectively constitute (trans)national sociologies. Building on the reception and adaptation of Wallerstein’s world-systems theory, there seems to be a bias towards studies in the dominant, and less so in the dominated periphery, and here Sweden can form an anticipated case seemingly positioned between the global center and the global peripheries.

Adding on the assumed existence of national sociologies, the question arises of what kind of traditions – in terms of themes, styles, and methodology – can be mapped out in a smaller and less studied country, and of how it is expressed in the type of sociological knowledge produced in its various localities. This opens up an array of potential research problems touched upon in the literature, like whether there is a general favoring of either qualitative (as the UK) or quantitative (as in the US) methodologies, or a balance between both. This is why the research strategy proposed in this dissertation (and developed in Chapter 5) takes for granted that for each computational text analysis modeled on the global level of the corpus, the dissertation also ought to scrutinize the effect of time and place on said sociological knowledge.

Grounded in the tendencies found in the recent trends in the sociology of sociology under review in this chapter, a case of dissertations in sociology, which are defended at five different departments of sociology in Sweden over 40 years, has been proposed. The case will be represented by two corpora, one of abstracts and the other of full texts, that are scrutinized to see what it might suggest about the fragmentation, paradigmatic status, and social conditioning of sociological knowledge. With this sketch of a strategy for investigating sociological knowledge at hand, the next chapter (Chapter 3) will seek to develop a theory to investigate the bearings of the three theoretical models established in this chapter with computational text analysis. In the following methodology chapter (Chapter 4), the ontological and epistemological challenges associated with computational techniques will be presented and, thereafter, the research questions corresponding to the strategy developed in this section will be given (Chapter 5). (The sampling strategy utilized to identify the abstracts and full-text dissertations that are the data will be presented in Appendix B, while the techniques for studying the chosen sociological knowledge entities in their respective setting are presented in Appendices C-E.)
3 Sociological Knowledge in Theory

From the review of proceeding studies of sociological knowledge presented in the previous chapter, a way has been paved for an empirical study of prevailing understandings of sociology. Based on the suggested readings of the literature, a contribution can be made by launching a study of sociological publications produced at different localities to explore the paradigmatic status of sociology as a science and constitution in terms of themes, schools, and styles. First, such an exploration of potential paradigmatic candidates could generate deeper insights into the characteristics of sociological knowledge and how they relate to the experienced crisis of sociology in the shape of fragmentation. Secondly, a deep analysis of an unusual and multidimensional case situated in the semi-periphery of international sociology would enable an engagement with the more historically informed studies taking the social conditioning of knowledge as its basis.

With the aim of gaining insights into how sociological knowledge that has been and can be conceptualized can inform empirical studies, this chapter will revisit the pioneering books within the sociology of sociology. As was hinted at in the introductory chapter, this thread of work primarily focuses on Kuhn’s paradigm theory and through this perspective, notions like fragmentation and crisis become only two of many possible words for describing the unsteady nature of a multiple or pre-paradigmatic state. Indeed, while these early voices seem to agree that sociology has not established a paradigm, they appear divided on the problem of whether it is characterized by none, two, or many paradigmatic candidates. With this follows the question of whether the discipline should strive to establish a uniform paradigm to avoid self-destruction, or embrace the current state of things and explore the fruits of pluralism.

The purpose here is not to take a stand on these matters. Before the analyses are conducted, the dissertation seeks not to comment on sociology’s current state, nor to explain how we got here or suggest how it ought to develop. Indeed, following the later sociology of knowledge tradition (e.g., Bloor, 1976), the attempt of the empirical study is to give no epistemic privilege to any version or presumption of what sociological knowledge is. Further, it is not important whether these theories represent true or false beliefs of sociol-
ology and its surrounding social world. More modestly, the only function of this chapter is to delineate a theory that can enable an empirical study of sociological knowledge based on computational text analysis. To lower the expectations of readers invested in sociological theory as a field in its own right, it might be worth clarifying that the empirical analyses will take a mostly pragmatic approach to theory, which is reflected in the content presented in this chapter. Leaning on Robert Merton’s (1968, p. 465) approach to sociological theory, the following sections focus on describing a set of logically interrelated concepts that “constitute the definitions (or prescriptions) of what is to be observed” and are “the variables between which empirical relationships are to be sought”.

The chapter comprises six sections. First, it sets out to explore the early works in sociology of sociology to see whether they bring about insights that can inform a study of paradigms in sociology. Second, the chapter attempts to find a way to translate paradigm theory within the framework of computational text analysis. Third, a further operationalization of this framework is proposed by situating it closer to the discipline of sociology. With this follows three sections dedicated to delineating different ontological, epistemological, and methodological dichotomies in sociology that, hypothetically speaking, paradigms in the discipline could potentially call upon. The chapter ends with a short summary of the proposed theory that will guide the dissertation. The more technical aspects of how a study of sociological dichotomies can be achieved in practice are picked up in the methodology chapters succeeding this one.

Paradigm Theories of Sociology

Thomas S. Kuhn’s (1962) *The Structure of Scientific Revolutions* articulates a critical break in a positive notion common around the middle of the 20th century where science was seen as progressing in a linear direction toward the truth of things by cumulating facts. For Kuhn, scientists are not a single fellowship of truth-finders unshackled by the conditions of society. Rather, he holds that science is divided into groups of puzzle-solvers working in disjointed social worlds, where each field or discipline is contingent on its own rules that are exchanged over time. By stressing that the history of science is not a gradual development towards rationality but one characterized by revolutions where the collective puzzle scientists work is overthrown, paradigm theory articulates the arbitrariness of ‘what counts as (appropriate) knowledge’ within a ‘normal science’ – Kuhn’s term for an archetypical science like physics and other comparably successful natural sciences. Following Kuhn (1962: x, 37, 76), paradigms are “universally recognized scien-
tific achievements that for a time provide model problems and solutions to a community of practitioners”, that supply scientists with “conceptual and instrumental tools”, and that “so long as the tools of a paradigm supplies continue to prove capable of solving the problems it defines, science moves fastest and penetrates most deeply”. In turn, a crisis takes place when a discipline departs from normal science, where the scientists begin to reject its basic assumptions and open up for paradigms to fill the hole. If a candidate is agreed upon, and its main principles of scientific work are accepted by the leading majority of the discipline, a paradigm, and consequently the normal science, is re-established.

This reading of Kuhn’s theory mostly serves pedagogical purposes since it hardly covers all of the 21 definitions of the scientific paradigm circulating in his famous book (cf. Masterman, 1970). Indeed, as is often the case for the most influential works in science, there are as many interpretations as there are definitions of their main concepts. Hence, to add yet another impression to the pile seems like a fruitless endeavor. Nonetheless, after receiving multiple critical remarks on the vagueness of his main concept, paradigm, Kuhn updated his work in 1969 with a postscript. According to Kuhn’s postscript, the paradigm has both a more general and “sociological” meaning, encapsulating “the entire constellation of beliefs, values, techniques” shared by a scientific community and a more specific meaning of “exemplary past achievements”, which entails models and examples that “can replace explicit rules as a basis for the solution of the remaining puzzles of normal science” (cf. Kuhn, 1970, p. 175).

The reception of paradigm theory in sociology was closely intertwined with the emergence of the sociology of sociology. In the first bulk of work conducted in that field, the most influential book was undoubtedly Alvin W. Gouldner’s The Coming Crisis of Western Sociology from 1970. For Gouldner (1970, p. 412), the sociological tradition, in its conservative form, spans back to Ancient Greece in the form of Plato through a positivist stage and a Marxist period, but is fully crystalized in the American reception of the European classics of sociology performed by Talcott Parsons. Indeed, the author holds that “a single, organizing, intellectual center for the sociological community” arose from “Parsons’ system [which] was often a paradigm that gave coherence to the sociological community … [that] has been not so much exploded as picked apart and now is slowly expiring under the growing apathy of its audience” (Gouldner, 1970, p. 162). In Gouldner’s narrative (see also Gouldner, 1985), the gradual disassembly of a paradigm founded on positivistic ideals and the system-functionalist program is the root of the fragmentation process of sociological knowledge, which was also alluded to in the introductory chapter of this dissertation.
The winding road toward a paradigm suggested by Gouldner was to make social theory societally relevant again by performing what he refers to as reflexive sociology. Reflexive sociology here relies on the notion that a certain set of “domain assumptions” are unconsciously brought into social theories by the people constructing them, which is something sociologists ought to be reflective and open about (Gouldner, 1970, p. 46). Sociologically speaking, some domain assumptions resonate with some audiences and not others, like questions of good morality and the nature of man. The disagreement between the domain assumptions of the dominant theory and the domain assumptions of the people in the discipline is what caused the crisis in sociology, according to Gouldner (1970, p. 34).

Perhaps not surprisingly, it is sociology’s community of professionals rather than its epistemology that is the locus of the intradisciplinary paradigm discussion. Remembering Kuhn’s distinction, the initial translation of paradigm theory to sociology can be interpreted as ‘sociological’ rather than ‘specific’. Evidence for this reception can be read explicitly in the pioneering book A Sociology of Sociology when, for instance, its author Robert W. Friedrichs (1970, p. 325) states that “[Kuhn’s] exercise was neither the history nor the philosophy of science, it was rather in sociology, the sociology of science”. Diverging from Gouldner’s work, however, sociology is here conceptualized through two self-images, called the priestly mode, focused on order and systemic thinking, and the prophetic mode, drawn to see conflict and incite activism. The author makes the case that the history of modern sociology was dominated by a systemic viewpoint justifying social order, personified in Parsons, and is moving towards a critical one emphasizing social justice. Yet, instead of the idea that this mode will replace the previous, Friedrichs (1970, p. 328) states that “the prophetic and the priestly modes will continue in dialogue”, and envisions the emergence of a dialectic paradigm where social understanding leads to the discovery of social regularities.

The idea of sociology balancing between two paradigms or pre-paradigmatic candidates brought forward by Friedrichs can be found in later works. One famous example is Lepenies’ (1988) historical study of the emergence of European sociology as a struggle between literary scholars and scientists to define the object of the discipline, which will be dealt with more thoroughly later in this chapter. A clearer focus on the ‘specific’ aspect of the paradigm is found in arguably the most systematic application and development of paradigm theory to the case of sociology, namely George Ritzer’s Sociology: A Multiple Paradigm Science (1974). Following the work of Masterman (1970), which will be dealt with in detail below, sociology is here conceived to be “a multiple paradigm science” where “each of its paradigms is competing for hegemony within the discipline as a whole as well as within virtually every sub-area within sociology” (Ritzer, 1975, p. 158). Acknowledging the
problem of the manifold definitions of the paradigm in Kuhn’s original work (cf. Masterman, 1970), Ritzer (1975, p. 156) sets out “to develop a more adequate application of Kuhn’s ideas to contemporary sociology than those found in the literature”, specifically Friedrichs. For this task, Ritzer presents a definition of his own to scrutinize sociology:

A paradigm is a fundamental image of the subject matter within a science. It serves to define what should be studied, what questions should be asked, how they should be asked, and what rules should be followed in interpreting the answer obtained. The paradigm is the broadest unit of consensus within a science and serves to differentiate one scientific community (or sub-community) from another. It subsumes, defines and interrelates the exemplars, theories, methods, and instruments that exist within it. (Ritzer, 1975, p. 157)

Further, the author holds that predecessor thinkers have made “the error of mistaking theories for paradigms”, making the concept a “useless tool for analyzing the status of the discipline”, and argues that the paradigm must be represented by a scientific community or sub-community that has these four basic components: an exemplar, an image of the subject matter, theories, and methods/instruments (Ritzer, 1975, p. 158). The idea here is to offer a higher level of generality that can expose the intimate relationship between methodological and theoretical differences in the discipline.

By scrutinizing the history of ideas in sociology, Ritzer (1975, p. 158) delineates the four components of “the three basic sociological paradigms – social facts, social definitions, and social behavior”. The student of classical sociology probably instantly draws connections between social fact and Durkheim, and social definition and Weber. Yet, the two classics are here presented as exemplars bridging the social facts paradigm and social behavior paradigm. While not being “the most profound” of the theorists, Ritzer (1975, p. 165) claims that “Parsons was the only one to deal with all three paradigms”. Thus, as was the case with Gouldner, emphasis is put on Parsons’ synthesis. Ritzer’s (1975, pp. 163–164) point, however, is not only to theorize why the crisis has occurred but also to provide tools for further inquiry, where the “true test” of his theory is “whether it enables the reader to better understand the current status of sociology”.

While the problem for Ritzer was that his contemporaries were apt to put too much emphasis on theoretical tendencies and collectives of thinkers (e.g., system theory versus conflict theory), it appears from the review of the current literature that sociologists might tend to overemphasize methodological tendencies, which comes with the risk of simplifying sociology as instead being divided by collectives of technicians (e.g., quantitative methods versus
qualitative methods). Critical remarks stating that sociology cannot look for a paradigm on the grounds of methodology or subject matter, since it shares both with other sciences, have been raised during the 21st century (e.g., Brante, 2001). Yet, keeping an agnostic view, Ritzer’s notion of sociology as a science defined by multiple paradigms – each constituted by a combination of entangled subject matter images, exemplars, theories, and methods – will come in handy to expand the imagination of what characterizes sociological knowledge. What is missing in Ritzer’s work in order to apply it in a computational study, however, is prescriptive concepts or ‘what to look for in the data’, in the sense of Merton (1968).

Among the pioneers in the sociology of sociology, one of the most thorough investigations into sociology’s crisis and historical examples of its ‘divided nature’ can be found in Piotr Sztompka’s Sociological Dilemmas: Toward a Dialectic Paradigm (1979). This work will serve as a cornerstone for delineating prescriptive concepts when investigating potential evidence for one or multiple paradigms. Consequently, this particular book will be dealt with at length below. For now, it is important to note that Sztompka, in contrast to Ritzer, appears to underplay methodological differences and is, in this sense, more adjacent to Friedrichs’ train of thought. However, the strategy of Sztompka is not to argue for two or three paradigms but rather to present a conceptual model based on a multiplicity of dilemmas drawn from a wide variety of scenarios in the history of sociological thought. Two examples are how to deal with value judgments and whether sociology can explain social phenomena. Yet, instead of conceiving these disciplinary encounters as immanent to sociological knowledge, Sztompka uses them to set a path for philosophically resolving all dilemmas by establishing a single paradigm in a new reception of Marxism. As was the case in Gouldner’s version of what sociology ought to be, this prescriptive aspect of Sztompka’s work will not be taken into consideration.

Before summarizing this section, it is worth mentioning that not all sociologists of knowledge thought that sociology could be considered mature or commensurable enough to be compared along the lines of normal science, and consequently that it was not a proper object for studying paradigms (cf. Eckberg and Hill, 1979; Eriksson, 1997; Phillips, 1975; Steinmetz, 2004). However, as was suggested in the introductory chapter and shown in the literature review, the idea of the crisis in the Kuhnian sense and the quest for finding paradigms in sociology has lingered on. Thus, the question appears still to be highly relevant for the discipline. When turning to the potential learnings of the pioneering works in the field, emphasis will be put on the applicability of the paradigm theories to a computational study of textual representations of sociological knowledge.
In the books of the 1970s, there is a recurring story of the lost potential of the classics and even a lost paradigm presented in the work of Parsons. This narrative coincides with Kuhn’s original paradigm theory since a crisis ought to occur at times when the foundational presumptions of normal science are questioned. Further, sociology is conceived to be in a state where either two or multiple paradigms compete for sovereignty. In some cases, like Friedrichs, Gouldner, and Sztompka, it is clear what path, or to what candidate, sociology ought to follow to escape the crisis, namely Marxism. For the purpose of this study, Ritzer presents the most compelling case of sociological knowledge as constituted by multiple co-existing paradigms that might be exchanged over time. Further, Ritzer’s definition of the paradigm concept fits supposedly fragmented or pluralistic disciplines like sociology by emphasizing that paradigms are combinations of exemplars, images of the subject matter, theories, and methods/instruments. This might bring about a positive departure from the early literature’s focus on theory (cf. Ritzer, 1974) – e.g., systems theory versus critical theory – and the current literature’s focus on methodology (e.g., Erola et al., 2015; Payne et al., 2004; Schwemmer and Wieczorek, 2020) – e.g., qualitative methods versus quantitative methods.

To be applicable for empirical studies outside the conceptual and historical tradition, Ritzer’s model missed a key component stating how to delineate paradigms without a presupposed idea of the paradigm one is looking for. Instead of deciding on a set of theoretical or methodological schools, Sztompka demarcates six overarching dilemmas (each encapsulating several sub-dilemmas) that can be imagined in several different combinations. By drawing on these dilemmas, which seem to have existed throughout the history of sociology as we know it, Sztompka’s work holds the potential to look more agnostically at sociological knowledge and map out silhouettes of paradigms in data. Lastly, there are also things to be learned from Gouldner’s work, particularly the notion of ‘domain assumption’. To expand concerning the literature review presented in the previous chapter, domain assumptions can be read as a version of the social conditioning of sociological knowledge. For the study at hand, focusing on dissertations in sociology, we can imagine that sociology departments over time have developed diverging sets of domain assumptions (both explicit and implicit) such as that some theories, methods, or subject matters are seen as ‘proper sociology’ where others are not. This gives us two dimensions of sociological knowledge through the lens of paradigm theory. As a whole, it can be in a state ranging from a single paradigm, through two or multiple paradigms, to complete fragmentation. Yet, in addition, there can be institutional variations of sociological knowledge production that are not so much expressions of paradigms as they are ingrained domain assumptions (which can be detached from the current state of sociology). With this theory of sociological knowledge at
hand, let us move on to the problem of finding representations of paradigms in textual data with the help of computers.

**Tracing Paradigms with Computers**

In an analysis of *The Structure of Scientific Revolutions* presented to Kuhn by the likes of Karl Popper and Imre Lakatos, Margaret Masterman (1970, p. 65) – a previous philosophy student of Ludwig Wittgenstein and later pioneering computational linguist – finds that Kuhn presents 21 definitions of his paradigm in the book alone. Presumably, Masterman presents the most thorough and clear reading of Kuhn’s work thus far and will therefore serve as the key reference for operationalizing paradigm theory in this study. Worth mentioning is that Masterman (1970, pp. 65-66, 68) sorted these 21 definitions into three types of paradigms: sociological paradigms (e.g., “a universally recognized ... [or] a concrete scientific achievement”) “which can function when theory is not there”, as well as artifact paradigms (e.g., “an actual textbook or classical work” or “an analogy”) and metaphysical paradigms (e.g., “a map” or “a new way of seeing”) “which can be used as a puzzle-solving device”. After an attempt “to establish non-sociologically a Kuhnian paradigm” in order “to distinguish Kuhn’s paradigm view of science from some sociologically sophisticated hypothetico-deductive view”, Masterman (1970, pp. 76, 79) finds that what “distinguishes the two views from one another is that a puzzle-solving paradigm, unlike a puzzle-solving hypothetico-deductive system, has also got to be a concrete ‘way of seeing’”, referred to as a “crude analogy” with the following “logical characteristics”:

If a paradigm has got to have the property of concreteness, or ‘crudeness’, this means that it must either be, literally, a model; or, literally, a picture; or, literally, an analogy-drawing sequence of word-uses in natural language; or, some combination of these.

In any of these cases, I wish to say that a paradigm draws a ‘crude analogy’; and further to define a crude analogy as an analogy which has the following logical characteristics:

(a) a crude analogy is finite in extensibility

(b) it is incomparable with any other crude analogy

(c) it is extensible only by an inferential process of ‘replication’, which can be examined by using the computer-programming technique of ‘inexact matching’, but not by the normal methods of examining inference. (Masterman 1970, p. 79)
In other words, Masterman not only defines Kuhn’s paradigm as well as how it develops, but also suggests a way for studying it empirically that seems to have been neglected over time but that is picked up in this dissertation and discussed more thoroughly below. However, following Kuhn’s definition of “normal science” that is dictated by a single “paradigm, which gives a more central insight into the nature of the field, though restricting it and making research into it more rigid, esoteric, precise” (Masterman 1970, p. 74), it is clear that sociology does not live up to this standard. Thus, we return to the question of whether or not sociology is a “non-paradigm science” (which “is barely distinguishable, if at all, from ‘the philosophy of’ the relevant subject”), a “dual-paradigm science” (characterized by “two competing paradigms struggling for the mastery”), or a “multiple-paradigm science” (where “each sub-field as defined by its technique is so obviously more trivial and narrow than the field as defined by intuition and also the various operational definitions given by the techniques are so grossly discordant with one another, that discussion on fundamentals remains”) (Masterman 1970, pp. 73-74). As was suggested in the previous section, this very question is a key problem in the literature on the paradigmatic problem of sociology (i.e., its fragmentation or lack of commonality) and the most elevated interpretation was found in the work of Ritzer (1975, 1974). However, for a science as opposed to a philosophy, the most principal question perhaps is how this problem can be investigated empirically. Given the rapid development of computational methods and their application in the social sciences of the last decades, it is now possible to try to perform what Masterman and Ritzer could only imagine at the time.

Masterman (1970, p. 85) lays out two processes by which a paradigm can replicate itself, the second of which appears to fit a study of sociology best: the first process is “by mathematical or other rule-governed inference” and the second process is one through which sciences initially develop, referred to as “intuitive inference”, where ‘inference’ is literally any kind of permission to pass from one unit or sequence of units or state of affairs to another unit or sequence of units or state of affairs – but it is intuitive; it does not go by rules”. In Masterman’s terminology, echoing her training in analytical philosophy, the logic behind how crude paradigms become replicated is expressed as follows:

For if there is one operation more than another which is not intuitive, it is the entirely mechanizable operation of marking a replica, B', of an original, B. This replication cannot therefore be what Kuhn means. He much more means, that when B' is a replication of B, B' reproduces what, for some known purpose, P, are taken to be the main features of B. When a mathematical model, M, for instance, is ‘hung on to’ a crude paradigm, C, in
the manner which we have been describing, M, for some P, reproduces the main features of C. (Masterman, 1970, p. 85)

Thus, the idea is that in the case of intuitive inference, the crude replica is not a true replication of a preceding one, but rather the two correspond in some of its main features, where the selection of what these corresponding features are is determined by the succeeding in accordance with its purpose for presenting the crude replica. To translate Masterman’s interpretation of Kuhn’s paradigm theory to sociology, we could, for instance, imagine a new paper being published in the most prestigious outlet of a “sociology of” (i.e., a sub-field of the discipline). Whether or not this paper applies the most orthodox or unorthodox theory within this particular “sociology of”, it has reproduced the sub-field by getting permission to pass since the main features picked up by the paper are satisfactory for this “sociology of”. However, the reproduction of the sub-field is not mechanical since the main features are not stated (go by no rules), it is rather intuitive in the sense that what is taken to be the main features of the “sociology of” (of all potential combinations of main features that could be generated from the literature) serve one or many purposes (e.g., for taking the sub-field in an anticipated direction, for establishing a connection between two sub-fields, or, undramatically, for the author to be published and for the outlet to publish a paper).

Even though Masterman wrote her paper half a century ago, she finds in the computer science of her time potential methods to perform analyses of such a crude replication of main features, and explains a way that they could be developed in the future. The first is “mathematics of classification … i.e. the formalization of the process of finding Wittgensteinian families” and the second is a form of “pattern recognition … for making a digital computer make an ‘inexact match’ between two formulae which are highly similar to one another, but not quite the same” (Masterman 1970, pp. 85-86). The methods she implies echo machine learning techniques that have been developed decades after she was theorizing about these matters, which Masterman (1970, p. 87) to some extent foresaw when writing that the problem is that “replication, and controlling replications, is logically horrible” but “the artificial intelligence men have now thrown new light on it; and it is (I think) how Kuhn’s paradigm extends itself”.

A decade later, in 1980, Masterman (2005, p. 283) would return to Kuhn’s paradigm theory to show more practically how it could be analyzed, this time based on the simple definition of a paradigm as either “an analogically used artefact” or “an actual ‘crude analogy’, that is, an analogical figure of speech expressed in a string of words.” Again, it seems like sociology would fit better with the latter option, since it lacks artefacts in the style of, for instance, the mass-energy equivalence ($E = mc^2$) in theoretical physics. These
strings of words containing a “reiterative meaning-specification” is not as simple as “retaining only the spokes with ideas that occur”, Masterman (2005, p. 289) claims, when applied to “the full variegation of natural language”. Rather, sociologists draw crude analogies with words and by doing so are “proposing to open up a new much-to-be-talked-about semantic area, rather than just drawing upon old ones; and this we can only do by filling in already known positions, in known patterns, in new ways” (Masterman 2005, p. 292). These crude analogies have to be extended in some fashion, for which Masterman sees two paths. The first, called “analogical pile-up”, is to establish “an analogy cluster” by “simply drawing on more synonyms” and, the second, referred to as developing the analogy, is reached by adding “a set of structured knowledge-glossary, and then using an inference-schema to pass to and fro between them” (Masterman 2005, pp. 292, 295). According to Masterman (2005, p. 295), the latter path generates “a verificational vehicle” and the former “tends to make the aspect attributions progressively more abstract and metaphysical”.

Sociologists have indeed constructed several glossaries for sociological concepts over the decades (cf. Pathak, 2000), as well as books that compare and link concepts of different traditions over time; for example, Sociological Theory by George Ritzer and Jeffrey Stepnisky (with a development spanning 30 years and 11 editions). However, the standard practice of presenting new sociological theory does not seem to be to make a detailed inference schema of how the proposed concepts relate to previous theories. Rather, extending a crude analogy by piling up synonyms to the underlying paradigmatic idea appears to be standard practice. Imagine, for instance, how many crude analogies there are for Comte’s concept of order in sociology – what Parsons (1968, pp. 89–91) called the “Hobbesian problem of order”, entailing social (i.e., supra-individual) containments on individuals – that build on the concept by simultaneously adding and subtracting components (e.g., institutions and collective representations in Durkheim; status and class in Weber; fields and capital in Bourdieu; interaction order in Goffman …). Establishing a knowledge glossary that includes an inference schema for how the concepts of the most famous sociological theories relate to one another is, however, a dissertation in itself. And here lies one of the main benefits of the nine prevailing sociological dichotomies that will be laid out in the sections below. They will, tentatively, represent recurring notions in sociology onto which sociologists pile up their crude analogies.

Thus, to fully operationalize the concepts of Masterman’s methodology for studying paradigms, we first need a series of textual representations of sociological knowledge that have all been granted permission to pass. In Masterman’s terms, this could be accounted for “sociologically” by selecting a corpus of doctoral dissertations as the empirical material since they are texts
conditioned to be generated within the discipline of sociology that institutionalized sociology (a sociology department at a university) has assessed as living up to the standards of a PhD in the discipline (permission to pass). By default, these acts of awarding a PhD in sociology, which in the Swedish context entails a published dissertation, are seen as intuitive inferences since they are not done according to any clear rules\textsuperscript{47} (e.g., there are no prefixed criteria for research problems, theories, methods, types of data, and so on). This dissertation proposes also that features are all words or word combinations occurring in the dissertations that will be henceforth analyzed (this is also the meaning of the term features in many computational methods utilized for analyzing text data). The main features, however, are those that recur throughout the corpora of dissertations defended at sociology departments in Sweden that this dissertation treats as its data. These features will be scrutinized in the light of establishing crude analogies – defined as “an analogy-drawing sequence of word-uses in natural language” (Masterman 1970, p. 79). With this procedure, then, it will be possible to map out to what extent the dissertations exercise analogical pile-up and whether their recurring crude analogies can be interpreted to form crude replicas of one or several paradigms in sociology. For the latter task, we will have to investigate how well succeeding dissertations correspond to preceding dissertations in general and in a given context (i.e., the consistency of the main features over time, and the crude analogies in particular periods and places).

Following the increasing facilitation of digitizing texts and the application of state-of-the-art computational text analysis methods to the operationalization of Kuhn’s paradigm theory laid out in the previous paragraph, it will be possible to revisit the discussion of sociology and its potential paradigm(s) with empirically-informed arguments. Conceptualized in the style of Masterman, the problem to investigate is whether sociology follows the pattern of a non-paradigm science, dual-paradigm science, or multiple-paradigm science. Before heading on to describing more technically how this will be done in practice, we will have to establish what can count as representations of crude analogies within sociological knowledge. Connecting to the previous section on the early paradigm theories of sociology, this will include returning to the fundamental dilemmas or, as it will here be conceived, dichotomies of sociology.

\textsuperscript{47} This is referring to “rules” in the factual meaning of the word, not “preconceptions”, “established viewpoints”, and so on (see p. 84 in Masterman 1970 for a longer discussion on the matter).
Delineating Crude Analogies through Dichotomies

The purpose of this section is to form a bridge between the two previous sections, the first dealing with the reception of the paradigm concept in sociology and the second discussing a way forward for theoretically-driven empirical studies of sociological knowledge through mapping crude analogies in text with the help of computers. This bridge will thereafter lead to three sections dealing specifically with three different types of potential crude analogies in sociology in the form of common theoretical oppositions in sociology. This theoretical move is inspired by a popular reference in the empirical studies of sociological knowledge reviewed in the previous chapter, namely Abbott’s (2001, p. 5) Chaos of Disciplines, where it is argued that “a defining characteristic of sociology” is that “the discipline is not very good at excluding things from itself”. Abbott argues that the justification of sociology as “a general social science” is not its scientific contributions, but that no form of knowledge about society is excluded from it, making the discipline “irremediably interstitial” (Abbott, 2001, p. 6). This creates a situation where all types of opposing sociologists coexist and try to reduce the discipline according to their particular image of what sociology ought to be. To make sense of this variety of positions within sociology, the discipline is distinguished by “a set of dichotomies” that “every graduate student learns” (Abbott, 2001, p. 10). Abbott argues that a dichotomy is a “fractal distinction”, meaning that it “is internally divided by the same distinction”, “that profoundly shapes our understanding of our own and other’s social science” and “measures our similarities and differences no matter how great or small those may be” (Abbott, 2001, p. 13).

Building on the work of Masterman (2005, 1970), these sociological dichotomies can be conceptualized as crude analogies that might be called upon to form paradigms in sociology. While different sociologists use different terms (and often interchangeably) to address these oppositions – such as distinctions (e.g., Abbott, 2001), dualisms (e.g., Fox and Alldred, 2018), and divisions (e.g., Azarian, 2021) – the most telling and pervading word appears to be that of dichotomies (e.g., Hess, 1990; Jenks, 1998; Pettinicchio, 2012; Stewart, 1958). The term dichotomy – coming from the Ancient Greek word διχότομος that is derived from διχότομος (dikhótomos), originally entailing “cutting in half” – has come to entail a “state of having a dual arrangement or order”. It is reasoned that the term dichotomy best resonates with the conceptual nature of these oppositions as well as the context of exegetic theory and adaption of philosophy in which they are discussed within sociology. Further, the term dichotomy is believed to capture the arbitrariness of

---

the opposition; while it might be an empirical fact it is, so to speak, constructed by societal relations within a certain frame of time and space. In comparison, while terms like dilemma and division seem too non-specific, the term dualism appears more fixated ontologically, analytically restrained, and, thus, less open for contradicting cases lying in between the oppositions. Thus, the term dichotomy, and this understanding of it, will be used in this dissertation.

The guiding strategy for delineating the dichotomies will not echo the technique of the literature review as presented in Chapter 2. Rather, the procedure can be described as a conceptual mapping of understandings of sociological knowledge that are understood to carry a certain longevity. The trustworthiness of this mapping is upheld not only by logic but also by staying as close as possible to the texts in question, which at times has entailed something of a tedious practice of stacking quotes. Nevertheless, the underlying premise of the following section is that these understandings of sociological knowledge tend to come in the form of theoretical oppositions that inform sociologists when speaking about the current state, as well as the past and future, of sociology as a knowledge-generating discipline. Amongst several other theorists, Sztompka’s (1979) work on the six sociological dilemmas will be a key source of inspiration for delineating diverging versions of sociological knowledge.

As has been alluded to, the dichotomies presented below are believed to represent crude analogies of sociology and an initial theoretical argument for how they can be studied empirically was given in the previous section. Nevertheless, it might be worth mentioning that many of the sociologists addressed in the last three sections – which formed the basis for delineating the traditional dichotomies of sociology – do not think that sociology is neatly divided into one or a few dichotomies. For instance, Sztompka, who once wrote Sociological Dilemmas (1979), has more recently expressed his dissatisfaction with “the almost obsessive theme of ‘paradoxes’, ‘ironies’, ‘dualisms’, or ‘dualities’” in sociological thought and even made a case against “exegetical theories” like his own 1979 work as “the frantic search ‘in a dark room, for a dark dog, which is not there’” (Sztompka, 2004, p. 263). Nevertheless, as has been deliberately repeated throughout this chapter, the argument put forward is that these formative dichotomies for sociology still appear to function as points of reference (and seemingly even sorting mechanisms) when the discipline tries to make sense of itself. The general tendency found in the literature review of the previous chapter is that authors are apt to address either a single dichotomy or multiple dichotomies side-by-side
without engaging in a deeper discussion of how they might interact (Abbott, 2001, is a clear exception here).49

Following the sociology of knowledge tradition, this chapter will not be involved with either the debate on the logical value of the dichotomies (whether the dichotomies are substantiated by reason or theoretically false), or the conversation of the ethical value of the dichotomies (whether the dichotomies are good or bad for sociology). Both of these questions are out of the scope of this dissertation.50 Thus, the reason why the dichotomies are here summoned is their long-lasting position within the disciplinary understanding of sociologists’ theorizing about sociology. For the sake of sustaining this assumption with textual evidence and avoiding accusations of misrepresentation, the chapter will include several direct quotations that might interfere with the flow of the chapter. As we will see, despite uncountable attempts to overcome the dichotomies henceforth presented, they linger on. Therefore, it seems safe to state that if they are not formative for the discipline’s understanding of itself, they are at least inert components of how sociologists reason about sociology.

To recap, the overarching idea, then, is that paradigms are constituted by crude analogies, which in this study of sociology is operationalized as a combination of a particular set of dichotomies that are repeatably found in a large representation of sociological knowledge (in this case, a corpora of dissertations). Accordingly, the following sections will deal with finding the most demonstrative dichotomies of sociology. The dichotomies are divided into three, which can all be related to Ritzer’s definition of the paradigm concept in sociology. In the first, the focus is on dichotomies stressing ontological dilemmas, particularly what the social world comprises. The ontological aspect of sociological knowledge connects to sociology’s image of its subject matter, which can be institutions, actions of individuals, and so on. The second set of dichotomies is primarily epistemological in the sense of what can count as proper sociological knowledge. This evokes questions related to objectivity and the involvement of values in research, as well as the relationship between empirical observation and theories, which connects with theory in Ritzer’s paradigm concept. The last section will be dedicated to methodological dichotomies, echoing the preferred methods of the para-
digms, and circles questions like whether sociology ought to lean on the humanities or the sciences, and use qualitative or quantitative techniques.

Before heading on, a last note on the division of the dichotomies might be fitting here. The dichotomies are, in themselves, not defined by clear-cut boundaries but rather tend to overlap and to be used interchangeably or be addressed in blocks. This has been noted many times before, for instance by Abbott (2001), who lets quantitative/qualitative serve as a master dichotomy that encapsulates other common dichotomies such as positivism/interpretation, realism/constructivism, and structure/culture. Indeed, Abbott argues that a process where emerging dichotomies replace older ones and are later reinvented is at the very heart of how sociology changes. Nevertheless, the choice of sorting the dichotomies into three serves both a pedagogical purpose, as it is easier to get an overview through categories, and a theoretical one, as the division connects with Ritzer’s paradigm concept that is deemed to be especially fitting for a study of sociological knowledge.

Ontological Dichotomies

In the history of philosophy, metaphysics deals with the nature of reality, and one of its branches, ontology, centers on the “study of being and the essence of things”.

Within sociology, ontology refers to the social reality of societies and their individuals. Following the genesis of the discipline, Auguste Comte famously coined the term sociologie in the early 19th century by merging the Latin word socius, which traditionally referred to “associate” or “companion” but came to mean “society”, and λόγια (“logia”), the old Greek word for “to tell” or “treats of (a certain subject)” that now refers to “a speaking, discourse, treatise, doctrine, theory, science”. For Comte (1853, p. 239), sociology is set to generate “a history without names of men, or even of people” that corresponds to Comte’s society imagined as a world where humans are ordered similarly to the physical world. However, since sociology emerged as “the science of society”, so have diverging thoughts on what this society is and how the people that inhabit it ought to be conceived.

In 1917, Émile Durkheim (1982, p. 248) argued that “the distinctive characteristic of human societies, and the proper subject of sociology” is “the ‘institutions’ … [that] take on substance as individuals succeed each other

---

without this succession destroying their continuity”. In contrast, Max Weber (1978, pp. 13, 15) wrote in 1922 that “these collectivities must be treated as solely the resultants and modes of organization of the particular acts of individual persons” and by this sociology can “accomplish something which is never attainable in the natural sciences, namely the subjective understanding of the action of the component individuals”. Durkheim (1982, p. 252) once responded to a critique of his work with: “I do not deny in any way that individual natures are the components of the social fact … [but] as they combine together to produce the social fact, [are] they are not transformed by the very fact of their combination. … Is the synthesis purely mechanical, or chemical?” Two centuries later, there seems to be no consensus to be found on the ontological problem of sociology, i.e., the nature of society and individuals as well as the properties of the entities that sociologists are supposed to study (e.g., Sztompka 1979, p. 241).

In this section, what is suggested to be the three of the most prevalent ontological dichotomies are described. These are conceived as interrelated ways that help us understand how societies and individuals are constituted through opposing poles. As holds for this and the two following sections, the choice of these specific dichotomies was made based on reading primarily theoretical literature establishing various conceptualizations of sociology and comparing the works to find recurring positions. In this sense, the chosen dichotomies are thought to capture basic commonalities – therefore specificities that only occur within a single text are purposely sorted out – of how sociology is conceived within theoretical works presented in the discipline from an ontological, an epistemological, and a methodological point of view. As will be the tendency throughout this chapter, the dichotomies dealt with in this section sometimes overlap each other, and are not argued to be exclusively ontological but can contain epistemological as well as methodological elements.

The question addressed by Durkheim and Weber is often referred to as the problem of “individual versus society” (e.g., Adams and Sydie, 2002, p. 327; Alexander et al., 1987, p. 1), which has been seen as “the common assumption of sociology’s founding fathers” (Smith, 2014, p. 14). This problem or assumption is captured within the dichotomy of holism/individualism (also known, but not the same, as holism/reductionism, holism/atomism, and collectivism/individualism). According to Sztompka (1979, p. 242), the former pole, holism, refers to “social wholes to be separate, supra-individual forms

---

53 Sztompka (1979, p. 242) prefers the term collectivism to holism. However, when looking through the sociological literature, holism appears to be more grounded in sociological theory and the one closest associated with Durkheim (cf. Zafirovski, 2000; Šubrt, 2019). The collectivism/individualism dichotomy seems instead to appear most often in the form of
of reality with specific properties and regularities of their own – something other and something more than an aggregate of individuals”, while the latter, individualism,\textsuperscript{54} sees society “simply as pluralities of individuals, larger or smaller sets of aggregates – nothing else and nothing more”. Another, perhaps common, way of phrasing it is that individualism (or, more critically, reductionism) holds that society is the sum of its parts – in Weber’s terms, the result of countless actions of individuals actors – and holism argues that society is more than the sum of its parts – in Durkheim’s terms, the result of a ‘chemical synthesis’\textsuperscript{55} that transforms the parts. Thus, while Durkheim is often understood as the personification of holism, so Weber is the archetype of individualism (cf. Sztopmka 1979, pp. 290, 293).

Worth noting is that Durkheim’s work has also been interpreted as a rejection of this very dichotomy (see Hughes et al., 2003). However, the same idea of trespassing holism/individualism has been presented, for instance, in readings of Gabriel Tarde (e.g., Law and Lybeck, 2019, p. 147), Durkheim’s main intellectual opponent in French classical sociology, which is often associated with taking on the pole of individualism vis-à-vis Durkheim’s holism (cf. Scott, 2007). In all dichotomies presented in this chapter, there are contrasting interpretations set out to argue against a dichotomy, yet somehow these dichotomies continue to be used as tools for sorting out theoretical positions in sociology. Again, this is how dichotomies work – they are abstract simplifications reached by juxtaposing two things for theoretical gain – and this is the point of addressing them in this dissertation.

Schematized in this simplified way, then, one might wonder if the dichotomy of holism/individualism is merely a thing of the past or if it is still applied in sociology. In a special issue in International Review of Sociology “devoted to ascertaining whether an international new and original sociological paradigm called ‘relational sociology’ can be identified today”, Riccardo Prandini (2015, pp. 1, 3) states that the increasingly globalized scientific system has opened a path for a set of epistemological and methodological advances placed under the heading relational sociology. According to Prandini (2015, p. 2), ”this ‘relational turn’ was the critique of the well-

\textsuperscript{54} Looking through the literature on Weber and individualism in sociological theory (see Udehn, 2001), one finds that Weber’s individualism is often addressed as methodological rather than ontological, as it is in Sztopmka’s view. This position appears to be valid since Weber, so to speak, comes with the whole package and presents an ontology, epistemology, and methodology in one sociological framework. Nevertheless, while some methodological aspects of individualism are touched upon here (as well as in the succeeding section on methodological dichotomies under another name, explanatory/interpretative), the focus in this section is on the ontological aspects of individualism.

\textsuperscript{55} An extended quote is presented above (Durkheim 1982, p. 252).
established ‘individualistic-collectivistic’ ontologies and methodologies that characterized sociology until the early 1970s” and “the old struggle between individualism and collectivism was overcome” by the new world order of the 21st century – ongoing transformations such as globalization, digitalization, mediazation, etc. – which caused a reformulation of traditional methodologies and epistemologies provided by traditional 20th-century sociology. This state of things opened up for a “relational turn” in sociology that Prandini describes as “a real breakthrough [that] is extremely important because it forces sociology to specify accurately the ontology of society and social relation” (Prandini, 2015, p. 13). Thus, following Prandini’s reading, there is an idea that this dichotomy has become obsolete over the last few decades. Nevertheless, with titles like Spencer is Dead, Long Live Spencer: Individualism, Holism, and the Problem of Norms (Zafirovski, 2000), and Individualism, Holism and the Central Dilemma of Sociological Theory (Šubrt, 2019) recurring in the 21st century, the dichotomy appears to be far from superseded.

While the traditional dichotomy of holism/individualism might primarily show up in discussions of (mostly classical) sociological theory, the issue addressed by it is picked up in sociological research by another and currently even more popular dichotomy, referred to as macro/micro. According to Jefferey C. Alexander and colleagues in the anthology The Micro-Macro Link (Alexander et al., 1987), the dichotomy of macro/micro encapsulates two modes in sociological theory that approach the problem of the level of reality that sociology ought to occupy itself with. Namely, whether social phenomena dealt with by sociologists exist on the macro level (the macroscopic), referring to whole institutions, societies, and cultures, or the micro level (the microscopic), referring to the self and everyday world of interacting individuals. Traditionally, the macro level refers to a society in its totality, and often involves comparative studies of entities on this scale, like states and world religions. In contrast, an analysis of the micro level always starts and ends with individuals – focusing on their doings, sayings, social characteristics, and so on.

In Foundations of Social Theory, James Coleman (1990, p. 702) presented what would become an infamous figure, a “Casual diagram for relating micro and macro levels” in the shape of a “boat” (or a “bathtub”) that has gained an almost mythical status in contemporary sociology. Here the macro level is conceptualized as a system of action that influences actors and their resources at the micro level, but as these actors and their resources change, so does the system of action. The solution to solving the micro-macro problem seems here to lie in emphasizing individualism, and similar attempts were presented both before and decades after Coleman’s boat. One example is the “radical micro sociology” presented by Randall Collins (1988, 1981),
which underlines what macro-sociology can learn from micro-sociology. The editors of the widely cited *The Micro-Macro Link* (Alexander et al., 1987, p. 1) discussed above, argue that *macro* versus *micro* is an analytical distinction, so “to link it to concrete dichotomies – such as ‘individual versus society’ or ‘action versus order’ – is fundamentally misplaced” and, thus, argue instead that we ought “to conceptualize the micro-macro theme as a distinction between different levels of empirical reality.”

Nevertheless, after all these theoretical attempts, the *macro/micro* dichotomy remains a part of sociology for sorting out different entities and still seems to be seen as a central issue of the discipline, which can be sensed in titles like *Modeling Micro-Macro Relationships: Problems and Solutions* (Opp, 2011) and *Micro-Macro Problem in Sociology: Towards a Solution* (Maigari, 2021). In present times, a third element, the *meso* level, is sometimes included to address entities figuring somewhere in between the other two (for a discussion, see Ahrene, 2021, pp. 3, 100), such as groups, organizations, and networks. Irrespective of the analytical value of the meso level, some have argued that it “only has heuristic capabilities in the interpretation of a particular social context if the other two levels are not neglected in the analysis” (Serpa and Ferreira, 2019, p. 120). Therefore, it becomes an empirical question to what extent the *meso* level as well as the *micro* and *macro* levels are addressed in the sociology dissertations.

If the *macro* level society is a super-individual entity, as in Durkheimian *holism*, it would make sense that it would be able to influence individuals living on the *micro* level, and these enduring social *structures* ought to be the focus of sociology. On the other hand, if the *macro* level society is merely an aggregate of its population living on the *micro* level, as in Weberian *individualism*, the role of the actions of these individuals, their so-called *agency*, ought to constitute sociology’s object. This problem leads us to the problem of social order and the dichotomy of *structure/agency* (also known as *structure/action* or *order/action*). In a more juxtaposed and extreme form, the former pole refers to “totally dependent objects, devoid of any initiative or control concerning their own activities, striving toward goals set for them, along strictly predetermined paths, guided by values and following norms imposed on them”, and the latter to “totally autonomous subjects, initiating, regulating, and controlling their activities towards goals they had selected […] guided by values and following norms of their own choosing” (Sztompka, 1979, p. 242).56

---

56 Sztompka (1979) does not use the dichotomy of *structure/agency* but that of *passivism/autonomism*, but in this dissertation, we will stick with the former dichotomy since it was the one that occurred most frequently in the literature addressed in this chapter.
Again, it has been argued that structure/agency (i.e., passivism/autonomism) are merely analytical distinctions. For instance, Sztopka (2014, p. 277) has propagated a “third sociology as opposed to both the sociology of action and sociology of structures, or better as merging both of them in the synthetic, more adequate approach to social reality”. This is because, according to him, “Social wholes and human individuals have only virtual existence, their separation and mutual opposition is the product of false, distorted imagination: common-sense illusions, and theoretical as well as meta-theoretical fallacies” (Sztopka 2014, pp. 273–274). Most sociologists would probably agree that society in reality lies somewhere in between autonomous individuals with agency and constraining social structures. Nevertheless, the dichotomy appears to serve as a point of reference when developing sociological theory (e.g., Crossley, 2022) as well as conducting sociological research (e.g., Williams, 2003) and the concepts that sociologists use in both domains should seek to mediate the two extremes (e.g., Hvinden and Halvorsen, 2018). In one of the most influential books on the matter, Sociological Theory by George Ritzer and Jeffrey Stepnisky (2022, p. 466), the attempts to integrate structure/agency and macro/micro constitute one of the “major schools” in “modern sociological theory”.

This section has been centered on the explanation of three major and interrelated ontological dichotomies – holism/individualism, structure/agency, and macro/micro – that are argued to be widespread in the discipline of sociology and tell us something about some prevailing understandings of sociological knowledge. With these at hand, we can now move on to how sociologists have reasoned about the epistemology of their discipline.

Epistemological Dichotomies

Traditionally, the word epistemology refers to the “theory of knowledge” – encapsulating questions such as what counts as knowledge and how can knowledge be attained? For millennia, humans have generated knowledge of societies and groups of individuals in the form of commonsense understandings or social lore, and even centuries before a discipline of sociology would emerge, branches of philosophy generated theories of how society should be conceptualized. Thus, a central epistemological problem for sociology has been what counts as sociological knowledge and delineating crite-

---

57 *Sociological Theory* was originally published in 1983 and is currently in its 11th edition.
ria for separating this knowledge from other forms of social knowledge (cf. Becker et al., 1952)?

To put it in normative terms, the underlying question has often addressed what makes sociology “better” than its extra-scientific competitors. Indeed, as early as the beginning of the 20th century, an article entitled Sociology and Plato’s Republic was published in the American Journal of Sociology where key figure Albion W. Smalls (1925) presents a categorization of two tendencies in sociology. The first is sociologists who tend “to call everybody a sociologist who has ever rationalized about social relations [...]” where Plato is Exhibit A in the sociological museum” and, in contrast, the second is sociologists who separate between “the subjective way” and “the objective way”, “the way of the speculator and the way of the scientist”, and “between dialectical social philosophers and sociologists”, where the “finding-mark of sociologists is the adoption of a genuinely objective method of research into human relations” (Smalls, 1925, p. 515). In other words, either sociology is understood as a tradition spanning to at least 375 BC (when Plato’s Republic was published) or it began about two millennia later with the institutionalization of the discipline at the turn of the 20th century and its adaptation of scientific or objective methods.

Smalls himself is a proponent of the latter, non-inclusive tendency and with this move wants to establish a distinction between “proper sociology” that adheres to objectivism and “not proper sociology” that is grounded in subjectivism. A century later, this dichotomy of objectivism/subjectivism is still a rooted notion in sociology. Traditionally, “the objectivists have argued that the scientific method requires publicly observable, replicable facts, and these are available only in the area of overt behavior”, whereas subjective phenomena “can be studied only indirectly through their connections with overt behavior, if at all” (Diesing, 1966, p. 124). In contrast, “the subjectivists have argued that the essential, unique characteristic of human behavior is its subjective meaningfulness”, and that sociologists “must understand action from the standpoint of the actor, as a process of defining the situation, evaluating alternatives in terms of goals, standards, and predictions, and choosing to act” (Diesing, 1966, p. 124). Despite multiple suggestions made by influential sociologists like Anthony Giddens and Pierre Bourdieu on how to overcome the subjectivist-objectivist divide and the fragmentation that preceded it, the dichotomy of objectivism/subjectivism seems to still be a popular topic in sociological theory discussions (Mouzelis, 2000, p. 741).

Besides the objectivism/subjectivism dichotomy, the problem of distinguishing knowledge produced by sociologists from knowledge generated by lay people or scholars schooled in other disciplines also comes in a slightly different form. According to Piotr Sztompka (1979, pp. 188–194), classical
sociologists like Émile Durkheim and Max Weber understood that the subject matter of sociological knowledge – conceptualized as “the inquiry about the nature of man, the essence of society, the course of human history, and the role of the individual in society and in history” (Sztompka, 1979, p. 129) – was no different from what was dealt with in commonsense thought and philosophy. However, the two latter streams of social thought were intertwined with “both practical and evaluative orientations” and therefore “the image of sociology as free of any practical functions and any normative content seemed necessary for its emancipation” (Sztompka, 1979, p. 130). In this reading of Durkheim and Weber, the classics can be said to position themselves at the fact pole in the traditional fact/value dichotomy (sharing connotations with neutralism/axiologism or detachment/bias), which can be found throughout the history of science. While more modern adaptations of this position can be found in the works of George Lundberg – a key inspiration for the institutionalization of sociology in Sweden after WWII (cf. Larsson 2001) – Sztompka (1979, pp. 195–200) has argued that there were other influential voices like Gunnar Myrdal, who is said to have pointed out the inevitable biases attached to sociological knowledge, and Alvin Gouldner, who proposed that the idea of value-free sociology is a myth.

With the rise of the notion of public sociology at the turn of the 21st century, suggesting that sociology is too professionalized and ought to become not only more relevant to the public but also that sociological knowledge ought to be co-produced with public actors (e.g., Burawoy, 2004, 2005), the fact/value dichotomy (here perhaps in an interpretation closer to that of cognitivism/activism) has once again moved to the forefront. As Karl Marx famously wrote in a note to Friedrich Engels: “Philosophers have hitherto only interpreted the world in various ways; the point is to change it.” As is shown by Anne Mesny (2009, p. 675), some critics of public sociology hold that sociological knowledge is “superior to nonsociologists’ knowledge” by claiming that the former knowledge type is true and neutral whereas the latter is false and value-laden. Again, as was the case when sociology emerged as a discipline, its distinguishing features are thought to be its abil-

59 Public sociology is strongly associated with the former president of the International Sociological Association and the American Sociological Association, Michael Burawoy, since the concept is developed thoroughly in his presidential address to the latter association in 2004, For Public Sociology (Burawoy, 2004). Burawoy’s theory, consisting of four types of sociological knowledge, is presented in a later section of this chapter.

60 Mesny (2009) has described four positions in the matter: that sociological knowledge is superior to layman knowledge, that the two are complementary forms of knowledge, that they are engaged in a circular relation (sociological knowledge builds on commonsense and informs commonsense), or that sociological knowledge and non-sociological knowledge are homologous.

61 Mesny’s version of the fact/value dichotomy is referred to as “False vs. True” and “Neutral vs. Value-laden” (Mesny 2009: 275-276).
ity to generate scientific knowledge from rigorous analyses, and its neutral mode of describing society as it is presented factually by empirical analysis rather than making normative value judgments about it. The critique of this position is of course nothing new. Among others, the work of Karl Marx has been utilized for stressing that sociology should be measured on its practical relevance and ability to change society rather than by its own cognitive goals and quest for the truth about society (Sztopka 1979, p. 131).

In his immensely popular book, *The Sociological Imagination*, C. Wright Mills (1959) takes a clear stance against the *objectivism* pole as well as the *fact* pole, stating that sociology is a moral and political act. However, in the book, several chapters address a third epistemological dichotomy dear to sociology, namely that of *theoretical/empirical*, this time in the shape of two alarming trends in (American) sociology of the 1950s. Mills saw sociologists striving to generate universal “Concepts” with a capital C, a position he refers to as “grand theory” and associates with the system-functional program of Talcott Parsons, entailing a logical analysis of society in its abstract totality. In contrast, he also identified sociologists obsessed with “Method” with a capital M, dedicated to an “abstracted empiricism” that generated specific empirical abstractions rather than explaining social reality. The *theoretical/empirical* dichotomy in sociology seems to echo classical divisions found within the history of philosophy, such as between rationalists and empiricists, inductive and deductive logic, and *a priori* and *a posteriori* knowledge.

One might think that contemporary sociology solved the problem of *theoretical/empirical* by formulating sociological theories based on empirical research. However, when reading James A. Davis’ (1994, p. 185) paper *What’s Wrong with Sociology?* we learn that “today there is no such thing as sociological theory if you mean empirical relationships that are comfortably predictable and general enough to turn up across more than one topic. […] We have been at it now for a century without much luck”. Indeed, the *theoretical/empirical* dichotomy seems to linger on in sociology. One more recent example centered around this very dichotomy can be found in an article published half a century after the Mills’ book was released. In it, Scott Lash (2009, p. 175) claims that from the extended trinity (3+1) of classical sociology – Durkheim, Weber, and Marx with the addition of Simmel – sociology has been conceptualized from an *a priori* and rationalist viewpoint found in epistemology by moving from “the question of how knowledge is possible to the question of how society is possible”, which was, according to the author, picked up by Parsons and has stuck in sociology ever since. While this “mode of reasoning” fitted the challenges facing sociologists at “the end of the nineteenth century”, Lash (2009, pp. 182, 185) claims, the 21st century brings about another “social reality of global flows, mobilities, and uncer-
tainties”, entailing that “sociologists here should not be a priori and rationalist, but a posteriori and empiricist”. Thus, it is still hard to find a consensus on the question of whether sociology is too theoretical, too empirical, or has found an equilibrium between the two poles.

In this section, we have temporarily left the ontological problem of defining sociology’s proper entities in order to dig deeper into its epistemological problem, conceived as the criteria for what distinguishes and what should count as sociological knowledge. This discussion helped to carve out three key dichotomies found to capture long-lasting epistemological queries within the discipline – objectivism/subjectivism, fact/value, and theoretical/empirical. As the informed reader has probably noticed and might feel the urge to critique, both the ontological and the epistemological dichotomies are interrelated with issues falling within the domain of methodology. Methodology will therefore serve as the sole locus in the next section.

Methodological Dichotomies

Compared to well-known disciplines within the sciences and the humanities, sociology is young. It has thus emerged in an intellectual landscape constituted by already institutionalized disciplines, meaning that proponents of sociology had to consider these other alternatives to carve out a space of its own. As we will see, the main questions seem to have been centered on what disciplines sociology ought to align itself with in terms of methodology. Historically, the term methodology was used to refer to a “branch of logic that shows how abstract logical principles are to be applied to the production of knowledge”, but, within the scientific domain, it most often refers to the system of methods or a general research strategy that involves some theoretical commitments (cf. Howell, 2013). It is a well-known story that sociologie was Auguste Comte’s second choice for naming his new discipline because physique sociale was already taken, and for him “ideas of Order and Progress are in Social Physics, as rigorously inseparable as the ideas of Organization and Life in Biology” (Comte 1853, p. 401). While some have argued that physics still serves as an ideal science for sociology (e.g., Zald, 1991, p. 176), it has been historically established that psychology (cf. Sztopmpka 1979), sociology’s social scientific relative, made up one of the most pressing issues for many sociologists figuring in the discipline’s classical era at the turn of the 20th century – like Charles Horton Cooley (cf. Jacobs, 2009), Georg Simmel (cf. Frisby, 1984), and Gustaf F. Steffén (cf. Ginnerskov,

---

2021). The question took on a methodological character, asking whether sociology ought to build upon the promising research program set forth by psychology or distinguish itself from it. One famous example is the debate between Gabriel Tarde and Émile Durkheim, where the former argued in favor of psychology and the latter against it (cf. Candea, 2010, pp. 25–27). While traces of this query can be found in the debate around social psychology – whether it is a part of sociology, psychology, or associated with both – it today seems rather obsolete. However, as will be shown in this section, it is clear that the problem of sociology’s methodology, particularly its scientific status vis-a-vis other disciplines, is not. Again, the task is to describe the set of dichotomies used most prominently by sociologists but in this case for the problem of methodology rather than epistemology.

Perhaps the most famous historical study of sociology’s puzzling journey vis-a-vis other disciplines is Wolf Lepenies’ (1988) Between Literature and Science: The Rise of Sociology. In the book, Lepenies shows how the dawn of sociology in early 19th-century England, France, and Germany was characterized by an intellectual struggle between literary and social science scholars with divergent understandings of how industrial society ought to be studied. According to Lepenies (1988, p. 1), this aspect of the origin of sociology generated two opposing modes: “a scientific orientation which has led it to ape the natural science and a hermeneutic attitude which has shifted the discipline towards the realm of literature.” In broader terms, Lepenies touches upon a dichotomy in sociology that we could intuitively refer to as science/humanities. The original book title in German – Die drei Kulturen: Soziologie zwischen Literatur und Wissenschaft – is perhaps more explanatory for Lepenies’ goal, to conceptualize sociology neither as part of the literary nor the scientific tradition, but rather as a third culture in between the two. This position has been heavily supported but also criticized. For instance, Raymond Boudon (2002) argues that Lepenies’ ‘sociology as third culture’ entails that the works of classical sociologists like Durkheim would encapsulate both an aesthetic and a scientific expression of sociology, which the former argues is false.63

Regardless of whether or not Lepenies’ view of sociology as a third culture is correct, his work accentuates the origins and development of the science/humanities dichotomy that appears to haunt sociology to this day. For instance, Mayer N. Zald (1991, p. 175) conceives sociology as a “Quasi-Science, Quasi-Humanities Discipline”. From Comte throughout the 20th century, Zald (1991, pp. 172, 180) argues that sociology has been obsessed

---

63 Boudon might have a point since Lepenies (1988) even problematizes the view of Auguste Comte as the mythical personification of scientific sociology and argues that he wanted to merge more humanistic and artistic aspects into sociology in the latter part of his life.
with the idea of “becoming a science” but should “tie to humanistic modes of thought” to develop further. According to the author (Zald 1991, pp. 179, 181), sociologists have put a good effort into developing “the scientific side of the enterprise”, which includes “our concern for explanation, empirical evidence, and scope of generalization”, but unfortunately “humanistic sociologists are more interested in justifying an interpretive mode than they are in exploring the implications of their humanistic orientation”, which includes taking “the humanistic disciplines seriously as sources for methods, for substantive questions, or for the organization of knowledge”. Zald here undermines what can be seen as a methodological dichotomy encapsulated within that of science/humanities (which according to the author is a simplification that reduces the latter pole), namely explanatory/interpretative (or in its more value-laden form, positivism/interpretivism), which is popularly used in the discipline to distinguish whether sociological analyses seek to generate scientific explanations or ought to settle for generating interpretations of meanings in the style of humanistic disciplines such as hermeneutics. 64 Indeed, for Zald (1991, pp. 176, 180, 184), the initial “commitment to science” was modeled on physics and meant a search “for a few fundamental concepts or elements related to each other in law like generalizations, through fundamental forces or mechanisms” but now entails a commitment “to rigorous criteria of explanation”, while “a humanistic orientation” has come to entail “drawing upon civilizations values and traditional modes of interpreting and understanding them” as well as “moral suasion and social criticism”. Therefore, a not-so-uncommon notion is to couple interpretivism with a rejection of positivism as in, for instance, Zald’s (1991, p. 180) description of Norma Denzin as articulating “the interpretivist, antipositivist model”.

Indeed, Jan Balon and Johan Holmwood (2019) have shown multiple thorough accounts from within the discipline where the “impossibility of sociology as a science” is delineated. Two leading examples are drawn from the 1960s and 1970s, a time when the discipline experienced huge growth, where critical voices committed to what Balon and Holmwood (2019, p. 340) refer to as “ethnomethodology” (basically the works of Harold Garfinkel and Harvey Sacks) and “post-Althusserian Marxism” (a category that includes the Anglo-Saxon reading of Michael Foucault). According to these scholars, the respective positions presented an alternative model of sociology based on a critique of its deficiency to live up to “sociological positivism”, but as time went by, both became “an approach that complements other qualitative research methodologies […] [and] absorbed to the very interpretivism

64 Hermeneutics is a theory and methodology of interpretation that has traditionally been utilized in biblical texts. See Hans-Georg Gadamer’s (1960) _Truth and Method_ for a pinnacle work on hermeneutics and what would become its modern adaptation called “philosophical hermeneutics”.
ostensibly under criticism” (Balon and Holmwood 2019, p. 344). Not only does the article suggest the prevalence of the critique of sociology as a science, but also stresses the power of the explanatory/interpretative (or positivism/interpretivism) methodological dichotomy, as well as hinting at a third one referred to as quantitative/qualitative, to which we will return soon.

One of the possible reasons why the explanatory/interpretative dichotomy seems to be popular in sociology is that the “verstehende Soziologie” developed by Max Weber is built on a modified version of the traditional German form of this conceptual pair, “Erklären und Verstehen”, where the first is connected to hermeneutic methodology and idiographic explanations and the second to natural scientific methodology and nomothetical explanations (Apel, 1982, p. 19). At the same time, Émile Durkheim has repeatedly been associated with carrying on the positivistic legacy set forth by Auguste Comte, which in the former’s version seeks to generate laws of society by describing social facts that exist outside the consciousness of the human subject being studied (Durkheim and Lukes, 1982). This viewpoint on the oppositions between the two classics of sociology has been picked up in theoretical works of sociology that distinguish a dichotomy between two forms of sociology based on explanations of social structures and interpretations of the subjective meaning of action (e.g., Brante, 1980). As one can imagine, to serve the aims of each pole one would have to conduct different types of research, which leads us to the third and last methodological dichotomy of this section, quantitative/qualitative.

In The Debate about Quantitative and Qualitative Research: A Question of Method or Epistemology? Alan Bryman (1984, p. 75) writes that for “the past fifteen years, the debate over the relative virtues of quantitative and qualitative methodology has gained considerable impetus.” After a deep analysis of the literature propagating each pole in the 1980s, Bryman found that there is a tendency to divide the sociological research process along the quantitative/qualitative dichotomy. He asked “whether it is possible to establish a clear symmetry between epistemological positions (e.g., phenomenology, positivism) and associated techniques of social research (e.g., participant observation, social survey)” (Bryman 1984, p. 75). After uncovering

65 While idiographic knowledge is based on descriptions and evidence of particular phenomena, as is often the case in, for instance, history, nomothetical knowledge is based on generalizations that are applicable to a class of particular phenomena that in turn can form the basis for formulating theories and laws (e.g., Robinson 2011).

66 On one side of the debate, Bryman (1984:77) found “a natural science, and in particular a positivist, approach to social phenomena” and “a preoccupation with operational definitions, objectivity, replicability, causality, and the like” where “social survey is typically seen as the preferred instrument of research”. On the other, there was “a commitment to seeing the social world from the point of view of the actor”, “preference for a contextual understanding”, and
some contradicting examples,\textsuperscript{67} he concluded that the “basic problem with this line of discussion […] is precisely that it is a convention” and there “is no necessary 1:1 relationship between methodology and technique in the practice of social research” (Bryman, 1984:89). Nearly four decades later, we find that the most popular methodological dichotomy in current sociology is undoubtedly that of \textit{quantitative/qualitative}. The overarching narrative is that of two methodological schools that do not mix (at least not in a fruitful manner), one based on research methods like statistical analyses of survey and questionnaire data and the other on narrative analyses of interview transcripts, field notes from participant observations, and/or documents. Indeed, sociologists of the 21\textsuperscript{st} century find themselves with specific journals for each of the two poles\textsuperscript{68} as well as a lacuna of ways to address the experienced separation of qualitative and quantitative sociology, such as “the ongoing ‘war’ between methodological camps” (Kelle 2005), “methodological pluralism” (Payne et al., 2004), “methodological separatism” (Erola et al., 2015), and “the methodological divide” (Schwemmer and Wieczorek, 2020). When it comes to the methodological dichotomy of \textit{quantitative/qualitative}, Andrew Abbott (2001) has probably taken its ontological consequences the furthest by presenting a chaos theory of sociological knowledge change where it serves as a “master dichotomy” for structuring the whole discipline. According to Abbott (2001, p. 28), there are “affiliations between various poles” and perhaps “the strongest of these affinities is what we might call the methodological manifold: an affiliation between four or five separate distinctions generally labeled by the distinction of qualitative versus quantitative”.\textsuperscript{69}

\begin{center}
\begin{tabular}{ll}
\textbf{Quantitative} & \textbf{versus} & \textbf{Qualitative} \\
POSITIVISM & & INTERPRETATION \\
ANALYSIS & & NARRATIVE \\
REALISM & & CONSTRUCTIONISM \\
SOCIAL STRUCTURE & & CULTURE
\end{tabular}
\end{center}

\textsuperscript{67} The examples presented by Bryman (1984:77) include studies based on social survey techniques that do not appear to be founded on positivism, and those grounded on participant observation that does not appear to be “commitment to seeing the social world from the point of view of the actor”.

\textsuperscript{68} For example, there is, on the one hand, a journal named \textit{Qualitative Sociology} that has been published since 1978 and, on the other hand, one called \textit{Sociological Methods and Research} published since 1972, which is presented as “a leading journal of quantitative research and methodology in the social sciences” (quote from https://journals.sagepub.com/home/smr [accessed 2022-11-09]).

\textsuperscript{69} In \textit{Chaos of Disciplines}, Abbott (2001, p. 3) lays out a theory of “self-similarity” based on the notions of “fractals” and “microcosm”, where a pattern is expressed in its totality as well as its constituting parts. This is supported by examples where sub-fields of sociology are divided along the same lines as the whole discipline. In this system, Abbott (2001, p. 28) makes the \textit{quantitative/qualitative} dichotomy a key case for “general affiliations among various poles of my fractal distinctions [that] are strong and enduring” and presents the following table:
While the six dichotomies falling within the ontological and the epistemological categories are all understood to prevail within sociology’s understanding of itself, the three methodological dichotomies addressed in this section – science/humanities, explanatory/interpretative, and quantitative/qualitative – appear to articulate some of the most principal problems contemporary sociologists are seemingly facing. Whether the methodological dichotomies and the problems they address are reasonable or not, there seems to be a convention of adding multiple diverging tendencies into a “master dichotomy” that can explain what sociologists are experiencing in their social world. For instance, the dichotomies addressed in this section are all present in what Jonathan H Turner (2016, p. 289) more recently referred to as “Sociology’s Big Divide”, which runs “between a discipline that seeks to be a science and one that does not”. According to the author, this divide has existed in sociology for over a century, was primarily uncovered by sociologists in the 1960s, and can now be found in divergent academic journals.\textsuperscript{70} Indeed, in line with the theoretical position put forth in this chapter, Turner (2016, p. 299) states that “even if these schisms are overdrawn and somewhat stereotypical, they affect how editors and reviewers see the intellectual world”. Indeed, the nine dichotomies presented in this chapter are argued to articulate some key aspects of how sociologists see their discipline, so with those at hand we can now move on to discuss how they can be studied in the empirical world.

Putting the Theoretical Framework into Practice

In this chapter, a theoretical framework has been formulated for studying sociological knowledge that focuses on mapping out signs of paradigms in texts with computers. The chapter began by exploring how the early contributors to the sociology of sociology literature conceptualized sociological

<table>
<thead>
<tr>
<th>INDIVIDUAL LEVEL</th>
<th>EMERGENT LEVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRANSCENDENT KNOWLEDGE</td>
<td>SITUATED KNOWLEDGE</td>
</tr>
</tbody>
</table>

(\textsuperscript{Abbott, 2001, p. 28})

Regarding the quantitative and qualitative poles, Abbott (2001, p. 11) suggests that within quantitative sociology there is a qualitative pole (e.g., cluster analysis) and a quantitative pole (e.g., regression). At the same time, within qualitative sociology, there is a quantitative pole (e.g., formal analyses of society based on sampling) as well as a qualitative pole (e.g., full interpretation). However, these “dichotomies within a dichotomy” are not fixed but change according to a process that sometimes appears dialectical, where distinctions are generated (thesis/antithesis) and one is declared the winner by often incorporating aspects of the loser (synthesis), but through complex patterns of remapping, the new formation (thesis) is met with a new challenger (antithesis) and so on.\textsuperscript{70} Turner (2016, p. 293) writes that the “\textsuperscript{1960s thus exposed the BIG DIVIDE in sociology that had always existed from sociology’s founding in the United States in 1880s and 1890s; and this divide continues to the present day.”
knowledge. In this literature, it was suggested that Ritzer’s (1975, p. 157) adaptation of the original paradigm theory to sociology was found valuable by providing a looser, and for sociology more fitting, definition of a paradigm as resting on “a fundamental image of the subject matter”. As there can be more than one fundamental image of the subject matter sociology ought to deal with, there can be parallel paradigms in the discipline.

Thereafter, a strategy for delineating paradigms in texts was initiated by discussing the work of Masterman (2005, 1970) and proposing that dissertations in sociology have been given permission to pass in the discipline. The idea of this strategy is to investigate whether the main features that occur in a corpus can be interpreted to construct crude analogies – defined as “an analogy-drawing sequence of word-uses in natural language” (Masterman, 1970, p. 79) – of sociology that are repeatedly reiterated in the dissertations (analogical pile-up). The focus of such an investigation is to evaluate whether one or several sets of recurring crude analogies can be interpreted as crude replicas of one or several paradigms in sociology.

A proposition for how crude analogies could be operationalized in the case at hand was found in Abbott’s (2001, pp. 10, 13) idea that sociology is characterized by “a set of dichotomies” that “every graduate student learns”, which “measures our similarities and differences” and “profoundly shapes our understanding of our own and other’s social science”. This led to an exploration to find the principal dichotomies of sociology by revisiting the theoretical side of the sociology of sociology literature. This exploration ended with presenting nine major dichotomies in three categories: ontology, epistemology, and methodology.

Building on the work of Sztompka (1979), it was proposed that the main ontological problem of sociology has been of society versus the individual. It was further suggested that different aspects of the prevailing understandings of this dilemma can be captured by three recurring dichotomies, namely: holism/individualism, structure/agency, and macro/micro. In terms of epistemology, the prevailing queries in the discipline are interpreted to be concerned mainly with the position and involvement of the sociologist when conducting research. It is suggested that three dichotomies capture the main strands of the prevailing understandings of what counts as knowledge in sociology, which are: objectivism/subjectivism, fact/value, and theoretical/empirical. Lastly, it was proposed that the principal problems contemporary sociologists face when addressing their discipline might be simplified into the classical problem of whether sociology ought to apply the ‘hard’ methodology of the (natural) sciences or the ‘soft’ methodology of the humanities. Three dichotomies were suggested to capture the most prevailing
understandings of how knowledge ought to be generated in sociology: science/humanities, explanatory/interpretative, and quantitative/qualitative.

The nine dichotomies presented in the paragraph above will serve the purpose of representing crude analogies that when combined can constitute crude replicas of paradigms in sociology. With a theoretical framework telling us what to look for in a corpus, we can now move on to establishing a methodological strategy for mapping out dichotomies with computational analyses of dissertations in sociology.
4 Methodology

Before this chapter unfolds, it seems necessary to remind the reader of a statement made in the introduction, where it was written that this dissertation will position itself within the growing field that is research conducted by social scientists who see the potential of integrating quantified text data (analyzed through computational tools) with thorough interpretations informed by, in our case, sociological theory. Although many social scientists have embraced the potential embedded in computational social sciences, some sociologists stress the limitations of both computational methods and digital data (e.g., Bonikowski and Nelson, 2022; Kang and Evans, 2020; Keuschnigg et al., 2018; Lindgren, 2019; Lindgren, 2020; Nelson, 2019, 2020; Pääkkönen and Ylikoski, 2020; Spirling and Rodríguez, 2019). Related to this is the reason why this chapter is called methodology (as opposed to simply methods) and that is that, among others, computational text analysis (or, to use a term more widely spread outside academia, text mining) forms a system of tools entangled in epistemological commitments, rather than a loosely gathered set of tools.

In this chapter, the methodology informing the studies that comprise this dissertation is laid out. The main methods for processing the data are taken from a social science appropriation of computer scientific and statistical tools often referred to as computational social science (CSS). Although computational tools are gaining momentum in sociology, these tools are still at this stage in time (i.e., at the end of 2023) rather uncommon in this discipline. This is why it is necessary to start this chapter by briefly describing the field, spelling out the rationale for utilizing them, and discussing the potential problems of borrowing tools manufactured in a discipline with epistemological commitments other than those found in the social sciences. At the backdrop of this dissertation is, therefore, the argument that the interpretive dimension of running computational algorithms on corpora of texts needs to be emphasized. This calls for a discussion of how a hermeneutical mode of doing CSS can be brought forth by leaning on, among others, Paul Feyerabend’s (1975) anarchistic theory of knowledge to trespass methodological lines of division in sociology and harness insights from both (see Lindgren, 2019, 2020) qualitative and quantitative methodologies.
Exactly how this position can be translated into concrete research practice is laid out by developing a methodological framework concerning the study at hand. In this case, the framework differentiates between the actual computational techniques for text mining and computational text analysis that the analytical chapters of this dissertation will deploy, and the ontological and epistemological ‘study design’ upon which this dissertation is built. The approach hereby employed can be summarized as an attempt to utilize a “computational cartography” (see also Lee and Martin, 2021) set to investigate whether common theoretical explanations of sociological knowledge (see the theory chapter) are sustained or retained by computational analyses of dissertations in sociology. This is why the intuition behind the computational methods applied are described at the beginning of each analytical chapter, while their technical details can be found in the appendices. Moreover, it is against this backdrop that one can make sense of the content of this chapter, which attempts to give readers unfamiliar with the methods used, insight into ongoing methodological discussions in sociology and computational social science about these methods, while also offering insight into how the corpora that will be analyzed has been generated. Thus, this chapter ends with two sections describing the data utilized in the computational text analyses that the analytical chapters are based on. The first section describes how the dissertations in sociology were sampled, as well as the process of generating two corpora fit for analysis. Building on this section, the second section describes the characteristics of these corpora, including discussions of the representations and limitations of each of the textual datasets that this dissertation relies on.

The Computational Imagination

The rise of accessible Big Data, and advances in computer-assisted techniques for analyzing societal data, have undoubtedly caught the attention of researchers in disciplines spanning from the natural sciences to the humanities (Edelmann et al., 2020). While social scientists have been attuned to different terms to label their practice depending on their academic affiliation – such as culturomics (e.g., Michel et al., 2011) or digital sociology (e.g., Lupton, 2015; Selwyn, 2019) – the most widespread and comprehensive heading appears to be the emerging field of computational social science (hereafter, CSS) (e.g., Lazer et al., 2009). Today, CSS encompasses interdisciplinary research societies, specialized journals, and even graduate programs. Through its “capabilities to collect and analyze data with an unprecedented breadth, depth, and scale”, CSS has been described as a force entailing “a paradigm shift in scientific research methods” (Chang et al., 2014, p.
Put in other words, “CSS has the potential to accomplish for sociology what the introduction of econometrics did for economics in the past half-century, i.e., to provide the relevant analytical tools and data needed to rigorously address the core questions of the discipline” (Keuschnigg et al., 2018, p. 8). However, the methodology is not ready-made but rather calls for a good dose of skepticism before being put into practice, or as some sociologists may say, a healthy dose of sociological imagination is in fact needed to not engage with these tools in both reflexive and theoretically-informed ways (cf. Bonikowski and Nelson, 2022).

One key problem addressed in CSS is that the accessibility of new digitalized content has been shadowed by carelessness in terms of reflecting on the data’s origin, shape, and breadth (e.g., Nelson, 2019). Following a critical review on epistemological debates within CSS presented by sociologists Törnberg and Törnberg (2018), the path to avoid entails the tendency to naturalize social phenomena and trust them to hold an ‘elevated understanding of social life’, while the path to take should reflect on how digital platforms manipulate the data they provide and are agents in shaping social behavior. Thus, researchers need to make reflexive considerations when employing these techniques. For instance, before reaching a conclusion on what type of people are expressing a specific type of cultural behavior on X (formerly known as Twitter), one should consider where tweets originate from (i.e., which parts of societies tend to tweet most often), and how tweets are framed by the power logic of this particular social media system.

For this study of dissertations in sociology, finding out the origin and shape of data is less of a problem since these texts are openly accessible at several public libraries – a sizable chunk of dissertations defended under the 21st century are even freely downloadable for all internet users – and are pre-structured in the sense that there are established structures in place that control who can publish a doctoral dissertation. In addition, each dissertation undergoes several stages of scrutiny (for insight into how PhDs in sociology are awarded in Sweden and information related to the history of sociology in this country, as well as the system of universities and university colleges that award these degrees, see Appendix A).

A second problem with CSS entails the application of computational tools that are deployed in a careless manner, which some have called a technical indolence (cf. Nelson, 2019). This typically refers to social scientists who are simply finding and applying specific transcripts of code out of mere convenience, overlooking both that the tool might be ill-fitted for the case at hand and the possibility that there may be more accurate tools that could instead be deployed. The underlying issue is that analysts do not make the effort to create an arsenal of tools and learn the technical underpinnings of
each, and their limitations in relation to different types of data (e.g., Spirling and Rodriguez, 2019). In other words, one needs sufficient knowledge of the data, the architecture of the methods (e.g., Lindgren, 2019), and the strengths and weaknesses of the algorithms that are used to perform different tasks. Without this skill set, social scientists run the risk of being not only ill-prepared for making informed analytical choices but also ignorant of how the chosen set of algorithms affects the input data and the output results (e.g., Denny and Spirling, 2018).

There is another problem associated with the computational techniques that this dissertation relies on, namely the idea that algorithms are somehow neutral and value-free (e.g., Törnberg and Törnberg, 2018). The uncritical appropriation of computational ‘results’ has been shown to extend beyond the research phase since the results can cause troublesome patterns of inequality in the broader society (e.g., O’Neil, 2016). In this dissertation’s case, the ‘results’ generated through these techniques could also lead to a skewed understanding of the context that is the backdrop for the data that this dissertation analyzes (i.e., abstracts for, and full-texts of, dissertations in sociology defended at Swedish universities awarding PhDs). This is why this dissertation has an appendix, Appendix A, solely dedicated to the Swedish context, and why the analytical chapters that follow this chapter will all include a series of analyses that take into account not only time (operationalized in the form of time periods), but also space (in this case operationalized as the universities where the dissertations in question have been defended).

Having offered some ideas regarding the uncritical appropriation of computationally generated ‘results’, it seems necessary to mention that the problem of these techniques’ so-called ‘neutrality’ can sometimes be regarded as a result of “the computational imagination”. This imagination here refers to the context in which these computational tools were manufactured and the epistemological assumptions that their manufacturers hold. Viewed as such, CSS methodology, in comparison to statistics, can be understood as a system of methods imported primarily from disciplines related to computer science, which have traditionally been remote from the social sciences. Thus, in simple terms, applying this system of tools in a sociology dissertation to analyze the abstracts, and the full-text versions, of PhD dissertations in sociology defended at Swedish universities, calls for a reflexive discussion of the social scientists’ understanding of what dissertations are and how to study them in a study designed with the kind of aims and research questions that inform this dissertation.

To delimitate the position of this study on these matters, it is necessary to demonstrate how computational techniques can be combined with an interpretivist mindset. By comparing the “workflows” (research design, data col-
lection, and analysis) of computer scientists and social scientists studying group processes, Allen and colleagues (2017, p. 33) construct a typology of disciplinary differences. At the risk of simplification, computer science aims at developing the performance of models based on cross-validations of automated methods applied to collectively shared and annotated types of sensor data. In contrast, social science engages in hypothesis testing or inductive theory building to develop a theory by coding human behavior and running quantitative or qualitative analyses based on surveys or recorded data sampled from an adequate number of groups. Following this typology, computational methods are manufactured to perform well at specific tasks on heterogeneous but specific kinds of data. They are not developed to be sensitive to what limitations the data hold about what is to be explained or how the method and results relate to theories of what we already know of the phenomena of study. Since these questions are fundamental for the social sciences, each researcher must engage reflexively with the particular computational tools s/he wants to utilize.

It is easy to understand how “the computational imagination” is easier to accept and, thus, can be beneficial for sociologists well-versed with statistical tools and mathematical reasoning in their quest for new ways to develop explanations for social facts. But what about interpretative sociologists operating with texts usually falling within the tradition of qualitative research? As has been suggested by several social scientists (e.g., Denny and Spirling, 2018; Nelson, 2019; Pääkkönen and Ylikoski, 2021), the interpreter is embedded in the whole research process when s/he utilizes computational techniques. For instance, in the case of computational text analysis or distant reading (cf. Moretti, 2000), the interpreter selects documents, considers what parts of the documents fall within the target of the analysis, and puts the documents through an encoding process – i.e., converts the texts into coded form by transposing and removing unmeaningful words – for enabling pattern recognition. Already at these first steps of the process, several choices on how to establish reasonable boundaries for the empirical object have to be made by the researcher.

As two sociologists in the field have put it, “access to greater volumes of data … has necessitated more theory, not less … because theory inevitably enters the computational research process at multiple steps” (Bonikowski and Nelson, 2022, p. 1471). Even if one were to bracket all epistemological assumptions building up the algorithms – e.g., the assumptions that natural languages have a ‘structure’, that ‘meaning’ can be computed, that less interference in the model leads to greater scientific objectivity, etc. – it is still the researcher’s conceptualizations of the social phenomena or empirical objects in focus that are the benchmark for determining the value and success of the machine for the analysis being conducted. However, this is not necessarily
only a problem, but can rather be seen as an opening for the social sciences to trespass on ingrained and unproductive methodological boundaries. Some sociologists of a more analytical tradition have even gone so far as to suggest that these “explorative methods of pattern recognition in unstructured texts” might even carry “the potential to overcome the qualitative/quantitative divide in the social sciences” (Keuschnigg et al., 2018, p. 8), which leads us to the next section.

Interpreters Operating Machines

In the introduction to his “anarchistic theory of knowledge”, Feyerabend (1975, p. 18) states that the notion of “a fixed method, or a fixed theory of rationality, rests on too naïve a view of man and his social surroundings” and when scrutinizing the history of science and its search for “‘objectivity,’ ‘truth,’ it will become clear that there is one principle … anything goes”. Feyerabend’s controversial idea is that scientific discovery has not been developed by, and, thus, ought not to submit to, the boundaries of what is currently regarded to be commonplace methodology in an academic discipline. Seen in the light of unprecedented developments of computational power, tools, and data, the social sciences are now, like the natural sciences before them (e.g., Grenander and Miller, 1998; Smith and Sutcliffe, 1996), faced with the question, if they dare to expand their imagination, of what social science can do, or cherish the truths of their deep-rooted methodological camps or tribes.

Within sociology, methodological segregation is most clearly expressed in the so-called “great divide” of qualitative and quantitative sociology, tangled up by different ideas of whether sociologists interpret and explain social phenomena. Inspired by Feyerabend’s (1975) attack “against method”, a few sociologists (Kang and Evans, 2020; Lindgren, 2019) engaged in CSS have initiated a new attempt to trespass the gap by not only bringing attention to the interpreter’s role in computational research but also by utilizing quantifiable methods to study qualitative material. This dissertation is an attempt to further crystallize the arguments for why sociologists should incorporate more computational methods in their research, and why computational social scientists ought to reflect more upon the qualitative dimension of their work, and can draw on the learnings of hermeneutics.

As mentioned above, relying on CSS might also be risky since one can easily fall into the trap that is “objectivism” if one blindly accepts the premises of computational methodology but, as will be shown, so too is following methodological conservatism, since a more “radical” interpretivism entails a
threat of “subjectivism” or, even worse, “relativism”. The balancing act that one ends up trying to manage when understanding these challenges can best be described in relation to each particular analysis, as will become apparent in the chapters that follow. Thus, the analytical chapters need first to explain how the techniques employed work, which is in and of itself a pedagogical challenge. These chapters need also to present the ‘results’ of computer-generated analyses by acknowledging that the models of the data do not just ‘show’ things but rather suggest patterns in the data that can be interpreted through theory. When attempting the latter, the chapters that follow will employ a language that aims to be neither purely objectivist nor purely subjectivist. Instead, these chapters will rely on language that takes a middle position on each side of the quantitative-qualitative gap.

Since the present study targets texts produced within an academic discipline, the potential value of utilizing computational methods will be highlighted through debates held among social scientists studying the meaning of things and those situated within the field of science and technology studies. Regarding the former, there are, on the one hand, researchers of culture arguing for the value of computational techniques for enhancing or even replacing traditional qualitative studies of texts. The argument is that these methods – in contrast to the traditional “hand-coder” methods – do not impose interpretations based on prior assumptions (e.g., Lee and Martin, 2015) and are, therefore, able to “fix” the subjective dimension of text analysis by bringing in objective measures of meaning (Goldstein, 2019). Although this route to objectivism may seem enticing to some, it would not only be illusory – the interpreter plays a principal part in computational text analyses (e.g., Denny and Spirling, 2018), even in unsupervised machine learning – but also ill-fated, since it would take sociology back almost a century in methodological struggles and breakthroughs in legitimizing qualitative research (e.g., Blumer, 1969). On the other hand, we find cultural sociologists criticizing the new computational approaches to textual analysis from the standpoint of the hermeneutic tradition. One of the more vocal and hostile examples is dispraising the work as a ‘primitive brand of positivism” adhering to an ‘Ideology of Lexical Units’ (Biernacki, 2015, p. 317) and stating that coding is a form of “Alchemy” based on “detaching meaning from its context” (Biernacki, 2014, p. 177). The two antithetic positions towards computational analyses presented within cultural sociology reproduce the wider standstill between qualitative and quantitative research in the discipline and, thus, do not take us further in our quest to break the unproductive order.

Turning to science and technology studies, Kang and Evans (2020) make a compelling case not only for the empirical evidence of the divide but also for trespassing it by revitalizing Feyerabend’s (1975) anarchist war “Against Method”. Through interpreting the rhetorical structure of word patterns gen-
erated by text mining algorithms run on article abstracts, the authors manage to show how leading science studies journals are attuned to an ontology and epistemology that is either qualitative or quantitative and almost solely interacting within their respective camps. To simplify, the qualitative camp explores and interacts with constructed and situated knowledge to generate new theory while the quantitative camp confirms existing theory by measuring the significance of patterns in metadata to examine the publication process (Kang and Evans, 2020, p. 935). Kang and Evans (2020, p. 930) stress that this separation is unfruitful and want to trespass it by “exploding the boundary between qualitative and quantitative studies of science”. The solution suggested is to utilize “powerful machine learning tools” to study “large-scale digital repositories of complex, qualitative scientific artifacts”, and such “complex data demand qualitative sensibilities to determine a research focus and how to interpret, theorize, and qualify insights” (Kang and Evans, 2020, p. 940). Thus, by disregarding the methodological tribalism that characterizes sociology, there are profits to be gained that trespass the sum of the work of both qualitative and qualitative sociology.

For the qualitative side, Pääkkönen and Ylikoski (2021, p. 7), suggest that the main benefit of integrating computational methods in research conducted from a hermeneutic perspective is “to control the interpretive process and to reduce suspicions that interpretations are the products of a researcher’s particular biased judgments” and “to provide the researcher with resources to show that one’s interpretation is not arbitrary”. This is done not so much by claiming that the techniques can show “the objective meaning”, as there are dozens of algorithms to choose from that will generate vastly diverse results. Rather, in comparison to studies driven solely by the interpreter’s “subjective meaning”, a computational methodology can enable the analytical steps taken by the researcher to become transparent for its audience and reproducible by its peers (see also Nelson, 2019). Further, computers can support actual hermeneutic research practice. The tools available within, for instance, the field of machine learning can aid text analysis by both exposing truly unexpected patterns in the data (unsupervised) and assessing if our qualitative-derived interpretations really hold (supervised) for the whole corpus (cf. Goldenstein and Poschmann, 2019; Nelson, 2020; Stewart et al., 2022).

In contrast to traditional quantitative epistemology based on testing strong theories, incorporating a hermeneutic sensitivity can enhance sociologists’ ability to conceptualize social phenomena. Indeed, “the enumeration of word quantities requires in-depth interpretation, which involves theorizing about distinct qualities” (Kang and Evans, 2020, p. 941). The key is not to reduce the interpreter’s role but to acknowledge who is steering the computer, which enables the notion of “mechanical objectivity” shadowing quantitative
research to be replaced by a form of “interactive objectivity” that draws on the benefits of both computational and hermeneutic research, as well as more accurately signifying what is going on in practice (Pääkkönen and Ylikoski, 2021, p. 30). Phrased differently, one could say that by utilizing qualitative insights “many of today’s computational opportunities may bring us closer than ever to realizing the vision of the classic sociologists who wanted to study social processes in terms of systems, and often on a macro-level” (Lindgren, 2019, p. 8).

It is, hopefully, now clear that computational tools can be valuable for sociologists but is the relation only one-sided? Sociologists have argued that by bringing sociology to the table, CSS can become more driven toward generating and advancing social theory (cf. Bonikowski and Nelson, 2022; Lindgren, 2020). This is ultimately why the aim of this dissertation71 was crafted as it has been, and why the discussion chapter will dwell on this proposal further. For now, all that is needed perhaps is a reminder that sociologists like Edelmann and colleagues (2020, p. 62) have claimed that sociology can influence CSS “to produce new theories of human behavior or elaborate on existing explanations of the social world” in front of “solution-oriented social science that aims to predict human behavior for practical purposes”. Since sociology is aimed at describing, interpreting, and explaining social phenomena with the hopes of increasing our understanding of “what is going on”, this implies not simply learning how to run a set of algorithms on freely available big data and publish raw outputs of the model. Indeed, we “should not confuse changes in the costs of methodology for discovery with the typically high expenses of achieving insights” (Chang et al., 2014, p. 78). Rather, like every other social scientist, the researcher utilizing computational methods must engage in thorough engagement with existing theory throughout the whole research process. In other words, “social theory is needed for considering both the data, the methods, the ethics, and the results of the research” (Lindgren, 2019, p. 4).

But besides placing higher demands on the analyst, a theory-based or sociological approach to CSS would also help the researcher to make informed choices in the analysis. In the sections above, it was argued that text mining techniques can take text analyses in a more “objective” direction but only if we, at the same time, acknowledge that the “subjective” interpreter taking part in every step of the process affects the outcome of the results. So, how should we make these choices? In their critical analysis of text preprocessing for unsupervised learning, Denny and Spirling (2018, p. 185) conclude that

71 The aim of this dissertation is to shed new light on the crisis of sociology by empirically scrutinizing prevailing disciplinary understandings of the state of its knowledge production through computational text analysis.
we should use theory “to choose potential preprocessing steps on the basis that the information this removes or preserves is reasonable for the application”. In other words, the theoretical underpinnings forming the composition of the study ought to guide what computational tools are chosen and how they are operated.

If researchers lean on theory to determine what to do and specify the choices they make – from the selection of corpus, through preprocessing, and all the way to the end of the analysis – their research will not only become more transparent (Nelson, 2019) but also transcend both the risk of idiosyncrasy shadowing purely humanistic interpretations (Pääkkönen and Ylikoski, 2021) and the risk of alchemy attached to purely computational coding without reflexivity (Biernacki, 2014). Through merging computational methods with hermeneutics, the study at hand wishes to reject these two extremes: the reactionary humanist who discards (or even mocks) the machine and the positivist scientist who wishes to detach (or even ignore) the human steering the machine. Thus, following Pääkkönen and Ylikoski (2021, p. 30), we should not ask “whether some formal method eliminates the need for interpretation in some parts of the analysis process but whether the method facilitates access to important information in text materials and enables systematic collective scrutiny of interpretations”.

Hence, the aspiration is to end up in a favorable middle position where interpreters take control of and operate machines in line with their theoretical commitments for the benefit of analyzing research problems. As noted, this is similar to Kang and Evan’s perspective but also akin to the related “hacking social science” perspective suggested by sociologist Lindgren (2019, p. 3), arguing in favor of a Feyerabendian anarchistic social science where the research question is in focus – e.g., combining classical sociological theory, like the work of Georg Simmel, with data science – instead of adhering to either the quantitative or qualitative camp by asking “which theoretical perspectives have been conventionally agreed to be compatible with one another”. In the next section, it is explained how this position is realized in relation to the study at hand.

Discovering and Measuring Patterns in Text Data

Following the thread running through this methodological chapter thus far, we have reached the question: How is an interpretivist or hermeneutical perspective to be incorporated into computational methodology, and vice versa, in concrete research practice? Specifically, because the context of the study is dissertations in sociology defended at Swedish universities, the overall
strategy for conducting the analyses that constitute the empirical inquiries in this study is inspired by the six principles for computer-assisted text analysis set out by the three leading computational social scientists Brandon M. Stewart, Justin Grimmer, and Margaret E. Roberts (2022). First, theory is the starting point for engaging in computational text analysis. Recalling the theory chapter, the focus is on delineating crude analogies of paradigms in the form of dichotomies. Thus, rather than assuming that a computer ‘can read the dissertations from nowhere’, it is decided beforehand that the data will be looked at through this theoretical perspective and the computer will be steered in that direction. Second, computers cannot replace humans in conducting text analyses. Rather, the computer augments the researcher’s ability to process (numerous and lengthy) dissertations but the results would be almost meaningless if they were not scrutinized by the interpreter.

Third, the social-scientific research process is iterative and cumulative. In this case, the analyses are the result of a process of going back and forth between discovering, theorizing, and validating to determine whether the suggested results hold when compared to previous studies of sociological knowledge. Fourth, the hope of computational text analysis is not to one day find a single universal model of language that fits all studies. Rather, the motivation for applying these methods here is to generate useful generalizations for answering the specific research problem set out and, hopefully, generate some insights for future studies of sociological knowledge. Thus, fifth, the best-suited method is not given a priori but is fundamentally task-dependent. As will be shown, the methods applied in each analytical chapter are the result of an approximation of what combination of computational tools might be the best to answer the particular research question(s) in focus. Lastly, and in line with the sixth principle, it is essential to constantly validate the results generated by machines. The analyses put forth in this dissertation are, on the one hand, assessed with the help of statistical measurements (see the appendices) and, on the other, corroborated qualitatively in relation to the reviewed studies and the theory at hand.

Building on the six principles presented by Stewart and colleagues (2022), an attempt will be made to combine “expert human knowledge and skills at interpretation with the processing power and pattern recognition brought by computers” within a single methodology (Nelson, 2020, p. 5). This is done by conceiving the computer-assisted text analyses as conducting cultural cartography, a perspective imagined by Monica Lee and John Levi Martin (2015). To perform cultural cartography in practice is to view plots generated from running statistical and machine learning algorithms on text data as maps depicting textual representations of cultural expression. Yet, maps are always an approximation of the empirical world rather than a “true” reflection of it. Thus, the interpreter has to keep in mind that this approximation
overemphasizes some aspects and parts of what is being studied and under-emphasizes others. In situ, the researcher has to make a myriad of choices before landing on a decent map or visualization of “what is going on” (e.g., Denny and Spirling, 2018; Lindgren, 2019; Pääkkönen and Ylikoski, 2021).

Besides selecting a good map, one has to keep in mind that language models do not hold the objective “meaning” of single words or whole texts and cannot replace human-based interpretations (e.g., Nelson, 2020; Stoltz and Taylor, 2021). Recalling Hans-Georg Gadamer’s (1975, p. 269) famous words, interpreting texts requires a “hermeneutically trained consciousness” and “involves neither ‘neutrality’ with respect to content nor the extinction of one’s self, but the foregrounding and appropriation of one’s fore-meanings and prejudices”. The goal set forth here is to combine these learnings from hermeneutics with the argument presented above that the interpreter ought to expose the main prejudices within the scientific field and use theory to “discipline” the machine, instead of letting its internal logic steer the study design. The proposed solution is what has been argued that sociology can bring to the CSS table – around which economists, political scientists, and psychologists already have started to get comfortable (Edelmann et al., 2020; Metzler et al., 2016; Nelson, 2020) – namely to make explicit the theory guiding the interpretation. This involves developing a “data/theory approach” by “advancing our perspective on both theory and methods in parallel” (Lindgren, 2019, p. 6). This can mean utilizing computational tools “to develop new theories of the social terra incognita created by the rise of digital technology” as well as “to develop new ways of creating theory itself” (Edelmann et al., 2020, p. 74). In this case, this involves “using new types of data and methods to revisit old sociological questions that were once thought impossible to study” (Edelmann et al., 2020, pp. 73–74).

Thus, the cultural cartography performed here will start by engaging itself with claims postulated in the most prevalent theories in the literature to see how they play out concerning one’s empirical material (cf. Kang and Evans, 2020). In the present case of sociology dissertations, this entails drawing out claims made in the sociology of sociology – like the notions of fragmentation and paradigms spelled out in the previous chapters – to formulate research questions that guide what to look for when trying to detect patterns by running unsupervised machine learning algorithms on the digitalized corpora explained in the sections below. The idea is that, in performing these analyses, new arrangements will be exposed in the data. Even though automated text classification – when compared to traditional content analysis – is fast, and can generate novel patterns, there is the risk of hazardous interpretations, particularly when the data is unforeseeable (see Kim, 2021).
Thus, instead of blindly accepting the maps of the empirical material generated by the machine, the position taken here is that “computer-assisted methods can effectively complement traditional human approaches to coding complex and multifaceted concepts in the specialized domain of sociology” (Nelson et al., 2018, p. 226). Thus, following the idea of “augmented reading” suggested by Stewart and colleagues (2022), this will entail the “utilization of text-mining tools [which] facilitate a formal in-depth analysis” by “enabling first interpretations that in turn may provide a stronger focus for subsequent close reading” (Goldenstein and Poschmann, 2019, p. 84). In practice, deep reading of ‘typical cases’ of the corpora is added to hand-code the data. For instance, the analysis of linguistic styles presented in a later chapter combines tools from computational linguistics to generate maps that are interpreted and given a deeper meaning through the traditional combination of deep reading and hand-coding. This reflects what has fairly recently been called “Pattern Refinement using Guided Deep Reading” or “Hypothesis Refinement Using Human-centered Interpretation” (Nelson, 2020, pp. 14, 25).

Therefore, the techniques constituting the cultural cartography will not be performed in isolation. Rather, echoing the iterative character of research in the social sciences (Stewart et al., 2022), the analyses conducted intentionally build on one another. This means that the ‘results’ from one study feed into the ways in which the next step unfolds (which is why this dissertation comprises four different empirical studies that build on one another). In following these steps, this dissertation aims to accomplish a more holistic interpretation of the corpus which exploits the potential embedded in text mining while relying on the theoretical framework that informs this dissertation. This will be a reflexive undertaking that will inevitably instigate interventions in the research process, where the analyses will have to be revisited and re-attuned in different forms of cross-examinations. The term cross-examination is here used in two senses: first, as statistical evaluations measuring the validity of the models from which the plots are generated, and second, as interpretative evaluations. The interpretative evaluations include visual interpretations of each plot that in a cumulative fashion are aimed at answering the theoretical problem as well as reading subsets of the corpora to see whether the computer-based patterns make sense in relation to the interpreter’s impressions of the texts and his or her knowledge of the empirical context. Put in other terms, the “interpretive portion translates the computational output into sociologically meaningful concepts to enable researchers to draw more abstract conclusions about the social world that produced the data” (Nelson, 2020, p. 30).

This section has sought to show how empirical patterns are discovered, measured, and evaluated through what has been described as a cultural car-
ography (cf. Lee and Martin, 2015; Stoltz and Taylor, 2021). In the version of the methodology presented here, the key idea has been to leverage the potential of computational methods while situating the interpreter at the center of the research process. The concrete steps taken to perform this cultural cartography are described in greater detail at the beginning of each analytical chapter, as well as in the technical appendices. Yet, to get an overview, the next chapter includes an overarching sketch of what analyses are included and how they build upon each other. Before moving on to this sketch, the corpora forming the empirical foundation of the study will be presented.

Sampling Dissertations and Generating Corpora

In this section, the sampling strategy undertaken in this dissertation is laid out. This involves delineating the boundaries for what in this study will count as a dissertation in sociology defended in Sweden. Due to the character of the data, this sampling includes the procedure of assembling texts into a unified and bounded format, which in fields like computational social science and digital humanities is referred to as a corpus (a body of text). Thus, the process of generating the two corpora forming the data on which the analyses will be conducted is given, namely a corpus based on abstracts for the dissertations and another constituted by the actual dissertation full-texts. For the quantitively inclined reader, all information about the character of the two corpora in terms of their characteristics and limitations can be found in Appendix B.

The empirical object for this study is dissertations in sociology, and the goal has been to enable an investigation of the social conditioning of knowledge by comparing several localities in one peripheral country. The chosen case is Sweden, where the state grants universities the right to award PhDs according to a system of disciplines and fields of research called ‘research subjects’ that are limited by the Swedish state (see Appendix A for a longer discussion). For analytical and pragmatic reasons, the definition of a Swedish dissertation in sociology applied here is fixed and narrow: only PhD dissertations defended in Sweden that are registered in the research subject “sociologii” will be included. One main analytical benefit is that these dissertations are de facto what was considered to be ‘sociological enough to pass as sociology’ according to an exam committee of experts in sociology and adjacent disciplines. Another is that there used to be five main localities in the form of departments of sociology (with a few additions in the late 2000s), so the effect of time and place on the production of sociological knowledge can be controlled. Since no Swedish bibliography of dissertations in sociology exists to this date, the pragmatic benefit of this approach is that this provides
a consistent heuristic for deciding what dissertations to include and exclude from the sample.

While this sampling strategy has its benefits, it also comes with the consequence of excluding a good number of ‘sociological dissertations’ defended in Sweden, each category dealing with its own specified field of research. To begin, there are a few research subjects with sociology in their title that are excluded from the sample, such as analytical sociology, sociology of education, and sociology of law. These research subjects share the history that they have established their own tradition of bringing a sociological perspective to another discipline and continue to be seated within that place in the Swedish university landscape. Even within the five main departments of sociology, there are an increasing number of ‘sociological dissertations’ that are defended in other research subjects that have been excluded from the sample, such as social anthropology, science and technology studies, and social demography. Further, one can only imagine the wide variety of ‘sociological dissertations’ that are defended throughout the country in other research subjects seated at other types of institutions, like departments of information technology and gender studies or research institutes devoted to demography and public health.

Guided by the definition given above, an investigation took place between January 2018 and August 2019 to gather information from digital and physical archives (in Uppsala and Lund) to establish a bibliography. Based on this work, it was found early on that the first PhD dissertation72 in sociology was defended in 1951. Since the study targets a period beginning more than 70 years ago, a great deal of time and effort was put into finding, delineating, and structuring a bibliography. A guiding finding was that five departments controlled the production of PhDs in sociology until 1999 (see Appendix A). Thus, the first step was to consult these departments, of which only one had a complete online catalogue of the dissertations defended at their university. The other departments’ digital lists and archives were often incomplete or included incorrectly coded dissertations that further complicated the work.73 Regarding ‘physical books’, none of the five departments had a copy of all their dissertations ordered in a systematic fashion, and the majority had only a fraction of the books on a more or less accessible shelf.

72 This refers to the old “doktorsgraden” and the “doktorsexamen” introduced in 1969. Thus “licentiatsexamen”, both in its pre-1969 and post-1980 version, is excluded.
73 Perhaps interesting for readers based in Sweden is the fact that the most straightforward in terms of sampling and data collection was dissertations defended at Umeå University, since this university has all their dissertations (1970-2019) available online – including the digitized full-texts.
The next step was to consult institutions offering official statistics, and university library archives that provided numbers and lists that, when compared, were contradictory. The case was often that the right variables, particularly the research subject, were not included or were coded differently according to local variations. Thus, it was necessary to check for “evidence” that the theses actually existed in physical or digital form or were not defended in another research subject (e.g., social anthropology and labor studies were at times coded as sociology). The reliability of the sources was assessed by comparing the presence and absence of each dissertation and potential differences in its ascribed qualities. Thus, in the majority of cases, multiple sources had to be compared to determine whether a dissertation should be included or not. This includes physical and digital archives, unofficial dissertation lists and physical books kept by the departments. When possible, official statistics on PhD degrees provided by Statistics Sweden were used for reference value.74

Thereafter, the dissertations were coded by the metadata received from the various sources (university, author names, age of authors, library categories, page numbers, correct title, publisher, publication series, etc.). This procedure of data collection has included hand-coding metadata that has been ongoing throughout the work of this PhD. In the analysis, two variables will be in focus and therefore require more attention: year and university. When it comes to the variable year, one source would give the year of the doctoral defense, another the year of publication for the dissertation, and a third the year of the degree. Given the fact that some dissertations were published years after the doctoral defense, this was regarded as an unfavorable source. A similar critique can be aimed at the year of the degree since some sociologists who passed the PhD defense would not immediately (if at all) choose to process their degree.75 Since the doctoral defense is conceived as the moment when the dissertation manuscript receives ‘the stamp of sociology’, the year of defense was chosen as the variable along which the dissertations are sorted by time. For the purpose of the analyses conducted here, a decision was made to stop sampling the dissertations that were defended after 2019.

---

74 The value of this source is unclear, since not all research students who defended a dissertation in sociology have applied for their licentiate or doctoral degree. According to oral sources, one reason might have been that those who defended a sociology dissertation and managed to stay within academia often got a position as a lecturer at the university. Thus, the official degree was not required to support their PhD status.

75 For the year of defense and the year of the degree, the two dates would seldom differ, and if they did it was only by a single year; for instance, when the defense was held at the end of the year and the degree was awarded by the beginning of the next due to delayed administration from either the university, the Swedish Higher Education Authority (UKÄ), or both.
With regard to finding a fitting variable for the dissertation’s place, there is one variable for the university where the PhD defense took place and another for the place where the dissertation was published, which refers to the publishing house (be it public or private). Again, it is assumed that the place awarding the PhD in sociology is what matters, not where the publishers are seated. However, concerning the dissertations’ locality, it is important to note that it only denotes that the defense was held at that university and that the PhD degree was granted through it. For instance, there are hundreds of cases where the PhD candidate spent almost all of their time at a university college that funded their studies but which did not have the legislated right to award the student a PhD degree. There might therefore be a discrepancy between the university branded on the dissertation’s front page, the university where the book was written, and the context where the empirical studies took place – not to mention research stays, conferences, and so on. Nevertheless, it was found that the best indicator was the university where the PhD defense took place because it is this institution that awarded ‘the stamp of sociology’.

With a bibliography consisting of 1,005 dissertations in sociology, coupled with metadata at hand, the next step was to begin to find and generate digital texts fit for analysis. It was found that digital abstracts began to appear in a systematic manner in the mid-2000s and digital versions of the dissertations, often referred to as full-texts, in the late-2010s. To target a period longer
than 20 years, as well as fill in the gaps (and there were many) of missing abstracts, all dissertations missing a digital full-text from 1951 until 2000 were ordered and brought from the university library archives.\textsuperscript{76} Figure 1 illustrates in a physical manner the labor put into carrying these 550 dissertations from the shelves of the library to the author’s small office.

The goal behind this effort was to find some way to digitize parts of the dissertations, like abstracts or summaries. A summary was only systematically included in the dissertations written in Swedish, so it was not possible to utilize. However, a manual evaluation of all dissertations found that many were coupled with a “spikblad”, an official document that includes details about the dissertation and the doctoral defense. It transpired that from the late-1970s onward, this document began to include an abstract of the dissertation. Consequently, a decision was made to generate a corpus from the 1980s to the 2019s consisting of abstracts. To serve this purpose, roughly 200 ’spikblad’ had to be scanned and converted into digital text with optical character recognition (OCR), performed in Python with the package OpenCV. This process was not as clear-cut as it may seem, so a good deal of effort had to be put into fixing all the mistakes made by the computer by comparing all digitized abstracts to the original abstracts. Finally, 815 abstracts were added to the bibliography (with metadata) of all the 871 dissertations that were defended between 1980-2019.

Turning to digital versions of the dissertations, between a third and half defended from 2000 to 2009 and the wide majority of those defended between 2010 and 2019 could be downloaded in full-text. To get an acceptable sample, the roughly 180 authors in the bibliography missing a full-text had to be contacted via email, from which about a quarter sent over a digital version of their dissertation. While digitized dissertations from the 2010s onwards come in a format more or less ready for analysis, the same cannot be said for the 2000s. In a few cases, these sociologists only had their dissertations on floppy disks or in seemingly archaic formats. These digital texts had to be converted between different formats and computer programs to end up with a streamlined corpus. Further, the OCR coding often had to be redone for a decent result. Thereafter, all full-texts were converted into a single generic format, namely txt. In the case of the hundreds of dissertations including Swedish letters, this was not without its problems. At last, based on the sociology dissertations defended between 2000 and 2019, a corpus of 380 full-texts coupled with metadata was formed.

\textsuperscript{76} Most of these books happened to be located in the archives beneath Uppsala University’s main library, Caroline Rediviva. I want to thank all the unknown librarians who helped me to unearth and transport all these dissertations.
Of these 380 full-texts, 166 were written in Swedish. Since word occurrences and their surrounding context words are in focus – semantics and not syntax – the goal was to include all data in the same models. This can be done with multi-language models, but the sample was far too small in this case. Thus, a decision was made to translate the dissertations written in Swedish into English. Since almost all of these dissertations are so-called monographs, with an average of 254 pages (compared to the 188-page average of the whole corpus), translation was a time-consuming task. All dissertations were translated with the help of a deep learning language model called DeepL, which has been found to do well in translating scientific text from a foreign language into English (e.g., Takakusagi et al., 2021). In a systematic evaluation of DeepL, comparing human-made and machine translations of Shakespeare’s plays, it was found that DeepL was ‘creative’ in the sense of conveying some ability to shift perspective in different paragraphs and had an accuracy and fluency rate of above 80% in the sample (Hu and Li, 2023). One major limitation of this machine translation, however, was that the translations were found to be too literal. To evaluate the performance of the model in translating the dissertations in sociology, random samples were taken from the Swedish and English full-text versions and compared qualitatively. While translation at times altered the syntax and shifted the emphasis in sentences, it was found that in all samples all the keywords were transferred to the English version and kept at a similar proximity to each other.

The journey to a dataset was, however, not over. The rough full-texts (full manuscripts) generated difficulties since, for instance, they come with reference lists, tables of content, and prefaces that are not the bulk of the text that this dissertation aimed to analyze. Also, some methods relied upon required that the dissertations be in a single XML file so that metadata could be added to the full-texts. Further, there were several issues with illegal Unicode characters that would mess up the analysis. Thus, a Python script was generated to cut out these sections and sort all full-texts into one file. The datasets (the corpora coupled with metadata) were then imported to R through a self-made function and turned into a corpus format with the package tm and tidy format with tidytext. The form of preprocessing for these two corpora, such as whether stop words are removed and how the words are stemmed or lemmatized, shifts from analysis to analysis and will therefore be discussed case-by-case in the analyses. The more quantitatively inclined reader will benefit from reading Appendix B, where the specificities of the two corpora, their coverage over time and place, and their limitations are given and discussed.

With two corpora to be analyzed, we can now move on to formulating the research questions that will guide this study, which includes an outline of the overarching analytical strategy for answering them. This takes us to the next chapter.
5 Research Questions

As was stated in the introductory chapter, this dissertation aims to shed new light on the crisis of sociology by empirically scrutinizing prevailing disciplinary understandings of the state of its knowledge production through computational text analysis. To meet this aim, the dissertation poses a series of specific research questions informed by the theoretical and methodological framework developed in previous chapters, and the tendencies found in the literature review. These read as follows:

1. *If fragmentation is understood as a historical process where the content of the knowledge produced in a discipline becomes increasingly disintegrated, do the corpora offer evidence that sociological knowledge is in a state of fragmentation?*

2. *Given that a paradigm denotes a crude replica constituted by reiterations of the same crude analogies in a discipline, do the corpora offer evidence that there are one, two, multiple, or no paradigms in sociology?*

3. *Following the notion that the social conditioning of knowledge is expressed in the time and place of its production, do the corpora offer evidence that sociological knowledge is socially conditioned?*

Computational text analyses on the corpora that constitute the empirical material (i.e., dissertation abstracts and full-text versions of dissertations) are the principal approach that will be used to answer these three research questions. These analyses are structured cumulatively and organized as four different studies that build on one another. The results of each analysis are reported in separate analytical chapters, which will present not only the target, data, main methods, and underlying assumptions of each study but also their relation to the research questions above, and the answers that each of these analyses gives us. Thus, as stated in the previous chapter, thorough accounts of the intuition behind the analyses will be given at the beginning of each analytical chapter, whereas the technical details of the algorithms applied and models generated can be found in the appendices.

The first analytical chapter, Chapter 6, aims to investigate the first research question by exploring – from the most open angle feasible and through the
longest period available – whether the content of the dissertations in sociolo-
gy is becoming increasingly fragmented or if it expresses some form of
‘core’ that remains over time. This is investigated by mapping out recurring
‘word constellations’ in the form of correlated word occurrences. This is
related to research question two, investigating whether there are any evident
signs of paradigms in sociology in the form of a crude replica of a certain
image of the discipline that are upheld by a set of crude analogies that recur
in the corpus. Following the assumption that such crude analogies are signi-
fied by repeated combinations of words and can be operationalized in the
form of sociological dichotomies, the analysis is based on computing word
occurrences and their correlations in the abstract corpus. For technical de-
tails, please see Appendix B. These word correlations are then modeled
through networks to examine the extent to which they might reflect recurrent
crude replicas. Thereafter, a clustering algorithm is added to evaluate to
what extent the overarching word network might be divided into larger
communities that constitute crude replicas of one or more paradigms. Fol-
lowing the theory of the conditioning of knowledge upon which research
question three has been generated, the next step is to see whether these crude
replicas are patterned according to the place and year when the dissertations
under study were produced (i.e., the year and the university where the doc-
toral defense took place).

Building on the analyses of the abstract corpus presented in Chapter 6, Chap-
ter 7 attempts to offer more definite answers to what this corpus suggests
about the three research questions. Again, this corpus is selected due to its
longevity and systematic form. The analyses in this chapter will address the
first research question; it will explore the supposed fragmentation of socio-
logical knowledge, conceptualized as a growing variation in crude analogies,
by operationalizing it as a process where the diversity in research topics
is enlarged. In simplified terms, it is hereby hypothesized that if the number
of topics is increasing over time, there is evidence for fragmentation, while if
it is decreasing, or approximately remains the same, there is not. Topics,
operationalized as prevalent combinations of specific words in the corpus,
are generated with structural topic modeling, an unsupervised machine learn-
ing algorithm developed for detecting latent topics and measuring the effect
of metadata on their distributions (see Appendix C for the technicalities of
the algorithms and models applied). Turning to the second research question,
the topics will then be interpreted to explore whether they echo any or sever-
al of the dichotomies of sociology developed in the theory chapter. Thereaf-ter, correlations between topics are explored to see whether different com-
munities of dichotomies emerge. Simultaneously, the analyses presented in
Chapter 7 will also investigate research question three by looking at the ex-
tent to which the research topics are conditioned by locality. Can topical
representations of paradigms be found and, if so, do they recur over time and are they associated with different universities?

In comparison to Chapter 6 and Chapter 7, which build on methods familiar to computational social science and have already been used in the sociology of sociology literature, Chapter 8 proposes a more exploratory methodology to approach all three research questions from a different angle. Drawing on work published within digital humanities and computational linguistics, this chapter approaches crude analogies from the angle of linguistic styles that capture repetitive patterns of seemingly meaningless common words. The idea here is that linguistic styles might or might not pick up a set of crude analogies through the way that the dissertations return to certain phrasings that include some types of phenomena and exclude others. While the former could suggest the formation of crude replicas of a paradigm, the latter would point in the direction of fragmentation. To find the repetitive patterns constituting a linguistic style, the method employed in Chapter 8 requires large passages of text and is therefore performed on the full-text corpus. The linguistic patterns will be modeled through bootstrapped consensus trees of the most frequent words used in the full-texts, synthesized in the form of a network. Based on the structure of this network, linguistic styles might be revealed by delineating communities of full-texts that tend to use similar types of common words. The technical details of the algorithms and techniques applied in this chapter can be found in Appendix D. The features of the linguistic styles will then be explored by, on the one hand, a contrastive analysis that sorts out the 20 most specific tokens of each style’s most frequently used words that are most specific for each style and, on the other hand, close readings of 25 dissertations (the five most characteristic of each style based on their position in the network). Again, the social conditioning of knowledge will be considered by investigating how the linguistic styles are patterned in terms of time and place.

To further explore the interpretation suggested by the results of Chapter 8, Chapter 9 investigates an alternative take on the meaning of the paradigm and fragmentation. This involves applying methods that have not yet been deployed in the sociology of sociology to approach the research questions from another viewpoint. Perhaps the crisis of sociology is not a problem founded on whether or not the corpora express too large a variety of concepts. Given that “the meaning of a word is in its use in the language” (cf. Wittgenstein, 2009, p. 25e), the foundation of the problem might rather be sought in whether there is diversity in the meaning being attached to the foundational concepts of the discipline. Remembering Ritzer’s (1975, p. 157) formulation of a paradigm in sociology as resting on “a fundamental image of the subject matter”, the approach is based on finding whether divergent meanings of the fundamental image of sociology’s subject matter are
articulated in the full-text corpus. This is done by modeling how the dissertations use the concepts of sociology, society (the main research object of sociology), and the sociological dichotomies presented in the theory chapter. The technique applied in this chapter is a neural network model that generates vectors from words that can hold information about the context in which words are used, often referred to as word embeddings. Word embeddings are based on the hypothesis that natural languages are structured according to distributional semantics, which holds that words can be assigned their meaning by the company of words they occur next to. In practice, this necessitates a multitude of sentences in which the concepts are brought into play since data on the context words appearing next to the chosen target words will be needed. This is why this chapter relies on the full-text corpus to model the various meanings given to these fundamental concepts (see Appendix E for the specs of this model and the methods applied to analyze it). Building on the results of Chapter 7 and targeting the overarching problem of the conditioning of sociological knowledge stated in research question three, the next step is to unravel whether there are temporal, local, and stylistic versions of how these concepts are used (i.e., different distributional meanings).

With these three research questions as well as an overarching analytical strategy for answering them at hand, we can move on to the presentation of the results generated from the computational text analyses of dissertations in sociology. Recalling the methodological position that was developed in the previous chapter, the studies make up a practical effort to engage with computational methods from an interpretive point of view. This involves presenting visualizations of the results that are deciphered by detailed interpretations, something that might be an unorthodox application of computational analysis. Thus, one risk is that a reader well-versed in quantitative research might think that the analyses are too meticulous and lengthy, while a reader well-versed in qualitative research could, in contrast, consider the interpretations to be too narrow or internalist and lack sufficient context. If this were to be the case, one could be tempted to reiterate one of the conclusions that sociologists often draw when it comes to characterizing their discipline, namely that there is a methodological divide. In this case, such a divide would materialize in generating different expectations on the reading and appraisal of sociological texts since they are meant to appeal to different readerships.
At this point of the dissertation, an empirical investigation of the three research questions posed is initiated that will span over four analytical chapters. The present analytical chapter will examine all three research questions by exploring the corpus of 815 abstracts to dissertations defended between 1980 and 2019 from the angle of ‘word constellations’, which refers to combinations of commonly used terms within the frame of word correlation networks. Echoing what was discussed in the method chapter of this dissertation, a basic assumption within computational text analysis is the *distributional hypothesis* stating that neighboring words in texts are part of similar semantic arrangements. Thus, the main idea is that by computing what words correlate (repeatably co-occur in the sentences constituting the abstracts) and which do not, we can get a sense of the semantic arrangements that are characteristic for the abstract corpus. If we then add a clustering algorithm to a word correlation network, it will be possible to map out overarching word constellations.

The first research question, which deals with the fragmentation of sociological knowledge, is examined by looking into whether or not there are recurring word constellations that might be interpreted to represent sociological knowledge and therefore speak for or against fragmentation. However, without an informed theoretical perspective – as well as prerequisite knowledge about the specific genre of social science dissertations that this corpus of abstracts falls within – these word constellations would presumably be rather meaningless for increasing our understanding of sociological knowledge. Thus, the analytical strategy set out for investigating the second research question is to map out to what extent they can be interpreted along the

---

77 Worth noting is that to map out the word constellations properly, we need to perform a set of preprocessing steps. This includes removing the most frequently used words – such as between adverbs, conjunctions, and prepositions – from the corpus during the preprocessing stage since they are in most cases not meaningful for social science analyses.

78 The first research question reads: *If fragmentation is understood as a historical process where the content of the knowledge produced in a discipline becomes increasingly disintegrated, do the corpora offer evidence that sociological knowledge is in a state of fragmentation?*

79 The second research question reads: *Given that a paradigm denotes a crude replica constituted by reiterations of the same crude analogies in a discipline, do the corpora offer evidence that there are one, two, multiple, or no paradigms in sociology?*
lines of the nine sociological dichotomies presented in the theory chapter, which are theorized as potential traces of paradigms in a given corpus. More specifically, if a set of word constellations were found to repeatably pick up several sociological dichotomies, they could potentially be argued to articulate *crude replicas* of a paradigm. After these overarching word constellations are mapped out, the next step is to explore how they are distributed over time and place. The reason for this is to investigate the third research question,\(^8^0\) where time is here operationalized as the year a dissertation was defended and place as the university, specifically the department of sociology, where the defense took place.

The chapter is structured into four sections that build on each other. First, the words with the strongest correlations in the corpus will be presented to map out the most word combinations. In the second section, a clustering algorithm is added for the sake of mapping out communities of the words correlating throughout the corpus in the search for meaningful word constellations. Based on the same computation, the section that follows explores how the found word constellations relate to the aspects of time and place. Thereafter, in section four, the corpus is broken into sub-corpora to investigate what word constellations are most common at different times and localities, respectively. Lastly, a brief discussion is offered that is centered on what answers the key findings of the chapter are presenting to the two research questions.

**Word Constellations on the Global Level**

This section centers on exploring the most prevalent correlations between words in the corpus, based on the underlying notion that the word constellations they form will give an understanding of what general content the sociology dissertations are addressing. Again, the assumption is that the words neighboring a word in the text tell us something about its meaning. As is standard practice in most tasks of computational text analysis, so-called stop words had to be removed from the corpus to perform the computation.\(^8^1\)

---

\(^8^0\) The third research question reads: *Following the notion that the social conditioning of knowledge is expressed in the time and place of its production, do the corpora offer evidence that sociological knowledge is socially conditioned?*

\(^8^1\) Stop words refer to the most common and, for the sake of content analysis, least meaningful words in languages. In English, this would include words like “it”, “is”, “to”, “you”, and so on. If stop words were not removed in the present analysis, the correlations of these generic words (e.g., between “have” and “been”) would dominate and completely overshadow all potential words of interest. Thus, a dictionary containing the most common stop words in English was matched with the corpus to remove them from the abstracts.
Thereafter, the correlations were calculated based on how often the words co-occur in the text.\textsuperscript{82} Lastly, for the sake of interpretability, a filter was added so that only the most prevalent words and word correlations are included and displayed.\textsuperscript{83} For the pedagogical reason of trying to capture the most overarching word constellations within a single map, the results are visualized with the aid of network analysis. In this network, found in Figure 2, the nodes consist of words, and the edges connecting two nodes display the words’ correlation in the corpus.

In Figure 2, we find a network visualization of the words in the corpus that strongly correlate within the same sentences, which together constitute word constellations of various sizes. The figure also contains information on how many abstracts a word occurs in,\textsuperscript{84} indicated by the size of a node – spanning from small (\textasciitilde100) through medium (\textasciitilde200) to large (\textasciitilde300) – where the smaller sizes indicate that about one-fifth of the abstracts use the word and the larger sizes over a third. Further, we can also read how strongly words correlate with one another by the distinctiveness or thickness of the edges (lines between words) – where the thinner light gray edges indicate weaker correlations (\textasciitilde0.2) between tokens, while the thick black indicates the opposite (\textasciitilde0.6). Since the threshold is set quite high, all word correlations appearing in Figure 2 are noteworthy.\textsuperscript{85} Hypothetically, a network of densely connected and sizeable nodes would depict a prevalent word constellation in the corpus, while the opposite suggests that fewer abstracts have connected that specific set of words and thus it makes a weaker word constellation.

\textsuperscript{82} In this analysis, word co-occurrence is defined as a word W’s symmetrical context window of five; that is the five words positioned before (W^5, W^4, W^3, W^2, W^1) and the five words positioned after (W^1, W^2, W^3, W^4, W^5) the center word W.

\textsuperscript{83} After several tests with various parameters, words occurring in less than 10% of all abstracts (i.e., 82 out of 815) were subtracted. In the visualization, only words that correlate at least 20% of the times they occur in the corpus are included.

\textsuperscript{84} A word occurrence is here calculated as the sum of the number of abstracts in which that word occurs, where each abstract can only take the value of 0 or 1. For instance, if the word interviews is used 1 time in one abstract, 4 times in another abstract and 20 times in a third abstract, the total number of occurrences for interviews would be 3.

\textsuperscript{85} Only words that occur in at least 10% of the corpus (>82 abstracts) and that have a sufficient correlation with at least one other word (>0.2) are included.
Figure 2. Graph of word correlation networks generated modeled on the abstract corpus. The size of the nodes refers to how frequently the words occur across the abstracts constituting the corpus (minimum 82 abstracts), whereas the thickness of the edges indicates how strongly the words are correlated (0.2-0.6).

With these instructions for how Figure 2 can be read, we can inspect what overarching word constellations constitute it. To the left, we find the largest and most dense network of correlated words, consisting of 51 nodes, some of which represent the most frequent words in the graph (which is indicated by the size of these words’ nodes) – Sweden, empirical, research, and data. On the opposite right-hand side, lies the second most extensive network, in which seventeen words are situated. In comparison, the second network appears to be constituted by words occurring a little less frequently in the corpus. Finally, we have a set of smaller and more weakly connected networks.
that, in general, have lower numbers of word occurrences. At the bottom of this figure, there is one tetrad (i.e., a network of four nodes – tetrad) and five dyads (i.e., two-node networks), whereas there is one dyad at the top.

Presenting the words constituting each network in Figure 2 will be done to analyze how they relate to the discipline’s understanding of itself and, thus, be read through the lens of the sociological dichotomies presented in the theory chapter (i.e., science/humanities, explanatory/interpretative, quantitative/qualitative, objectivism_subjectivism, fact/value, empirical/theoretical, holism/individualism, structure/agency, and macro/micro). The most general interpretation put forward will be that the largest networks on the left-hand side of the graph primarily address empirical matters (empirical/theoretical), while the larger network at the top of the graph addresses theoretical matters (empirical/theoretical). The networks are further categorized according to poles of other dichotomies they are interpreted to articulate, such as the analytical level of society (micro/meso/macro) and whether they are expressing a more scientific or humanistic vocabulary (science/humanities).

Presenting the words constituting each network in Figure 2 will be done to analyze how they relate to the discipline’s understanding of itself and, thus, be read through the lens of the sociological dichotomies presented in the theory chapter (i.e., science/humanities, explanatory/interpretative, quantitative/qualitative, objectivism_subjectivism, fact/value, empirical/theoretical, holism/individualism, structure/agency, and macro/micro). The most general interpretation put forward will be that the largest network on the right-hand side of the graph primarily addresses empirical matters (empirical/theoretical), while the smaller networks address theoretical matters (empirical/theoretical). Informed by the theoretical underpinnings that guide these analyses, the former network has been divided into parts addressing qualitative and quantitative methods, respectively (qualitative/quantitative), the macro and micro levels of society (micro/macro), as well as capturing a more scientific vocabulary (science/humanities). In comparison, the smaller networks seem to pick up a more humanistic vocabulary as well as words alluding to both the micro level and the meso level.

If we look closer at the largest network, we can note that this part of the figure can be seen as comprised of three major parts: two smaller at the left and one larger at the right. The interpretation that will be put forward is that the separation of each part has a methodological as well as an ontological logic, where a dividing line between the poles is found in three words without a strong connotation to any of the sociological dichotomies, namely family, life, and gender. Beginning at the western side of this triad, there is a word constellation that could be interpreted as resonating with the micro level of society (micro/macro) and with empirical studies (empiri-
cal/theoretical) based on qualitative methods (qualitative/quantitative). This word constellation is connected to the largest network through two links. The first begins with life and everyday, which connects to experiences and experience. Put into one sentence, a good share of abstracts appears to address phenomena related to experiences and everyday life, which is interpreted as resonating with the micro level of society. The second link connects gender with the commonly used word interviews that further binds qualitative as well as strategies and material to the network. The latter forms a tetrad with two words referring to empirical analysis, the frequently occurring empirical and analysed, as well as the generic word consists. This word constellation is interpreted as echoing empirical studies based on a qualitative methodology. Worth noting is that qualitative is also linked to experiences, and the same goes for everyday and interviews, which might suggest an association between the mentioned methodological and ontological poles.

To the eastern side of the dividing line – again, made up by the triangle family, life and gender – are words that seem to allude to empirical research (empirical/theoretical) conducted with quantitative methods (qualitative/quantitative) and words seemingly drawing on a vocabulary related to the sciences (science/humanities). At the center of the largest network in the graph, we find a triad made up of survey, data, and results. The first node represents a quantitative method, the second a scientific term for empirical material, and the last node a quite generic word but one that is linked to research, analyses, differences, conditions, factors, influence, and effects, the sum of which suggest scientific rather than humanistic research, and measurement of some form of causality associated with quantitative rather than qualitative methodology. Worth noting is that while interviews is a token that correlates with both qualitative and material, which leads to empirical, the token survey strongly correlates with data as well as, to a lesser extent, results. In other words, qualitative studies might prefer to use the term empirical material and quantitative studies the term data.

By a single link, between one of the tokens addressed in the paragraph above, differences, and the token countries – two words that combined hint at some form of international comparison – there is a smaller constellation of words connected to the southern part of the largest network that indicates focus on the macro level of society (micro/macro). Besides the token countries, this includes sweden, public, national, and institutional, as well as words with a strong leaning towards the political sphere, namely political, policy, and welfare, where the two latter suggest a societal rather than individual focus. Another mentioned token, results, is connected to level that leads to levels; its neighbor factors, which was interpret-
ed to have a scientific connotation, is correlating with both structural — leading to structure and change — and individuals that further connects with individual. The latter set of words could be interpreted as leaning towards both a more individualistic (holism/individualism) and structural conception of society (structure/agency). Perhaps we can even speculate that this combination between the two sets of words refers to studying both individual and structural factors, as in quantitative methods able to generate multilevel models (qualitative/quantitative). In addition, results is correlating with words related to work and economy, starting with the node market that leads to the interconnected tokens labor and resources as well as to economic, which is linked with conditions and class. Thus, we can perhaps allow ourselves to assume a relationship between quantitative methods and the mentioned topics.

Surrounding the largest network, there are multiple small networks that are not as clearly associated with empirical epistemology, but seem to address theory (empirical/theoretical). Beginning with the far-right side of Figure 2, we find the second largest network of the figure, made up of two parts. To the north, there is a relatively dense selection of words that bring attention to sociology as such — sociology, sociological — and, particularly, sociological theory and knowledge — theory, theoretical, framework, knowledge, approach, developed. Further, we find that theoretical and framework co-occur with tokens such as discussion and discussed. While this set of words is alluding to theoretical sociology and, thus, the humanities, theory is of course a term used in both it and the sciences. However, the combinations of critical and theory as well as sociology and knowledge seem to give the northern part of this network a leaning towards that pole (science/humanities). It is also noteworthy that the sociology words are not part of the largest network and co-occur with theory words. This suggests that sociology might be referred to mostly in the context of theoretical discussions rather than empirical work and, thus, neither of the poles addressed in the largest network.

By following the edges southwards, we can find the token forms that is linked to form and organizational. The latter takes us to the node organization with links to organizations, leading to actors and over to local. These sets of words will be viewed through an ontological lens. One speculation for why words sharing the stem “organization” are isolated from the largest network is that they clearly exemplify the meso level (i.e., lie between the micro and macro level addressed in the largest network), but sustaining this would call for more fine-tuned computational text analysis tools. Further, the words local and context are interpreted as leaning slightly more towards the micro level (micro/macro) and the usage of the word actors, that is neighbored by local, is coupled with connotation to a society of individu-
als (holism/individualism) with agency, rather one determined solely by structures (structure/agency).

At the bottom of Figure 2 we find a tetrad which comprises the words practice, practices, interaction, and understanding. The first three words are interpreted as concepts commonly used in micro sociology (micro/intermediate), whereas the word understanding, in the context of its only correlation with the word interaction, is understood to suggest an interpretivist methodology rather than an explanatory one (explanatory/interpretative). Next to the tetrad, there is one dyad constituted by discourse and constructed and another of understood and relation, which are interpreted to echo interpretivism (explanatory/interpretative) as well as the humanities (science/humanities). This interpretation is based on the notion that theories of social constructionism and discourse analysis interpret society and tend to question the authority of objective science, which is often crudely accused of taking for granted that the social world is a truth external to the researcher subject.

Lastly, north of these smaller networks are three dyads that are harder to directly categorize in relation to one of the nine dichotomies that the theory chapter of this dissertation presented, and which are hereby being used to interpret the results of this first analysis. In spite of this, the word constellations characterizing these abstracts seem to resonate with the humanities (science/humanities). This seems to ring true for all three dyads – cultural and society; historical and development; defined and concepts – with their reference to culture, history, and (indirectly) philosophy, respectively, which are all disciplines that fall within the humanities rather than the sciences. When it comes to the last dyad – change and structure – it is interpreted to reference the structural aspects (structure/agency).

To wrap up the analysis performed in this section of word constellations in the form of a network based on correlations, the overarching pattern appears to be that many of the words with the strongest correlations can, more or less clearly, be associated to certain poles of the prevailing dichotomies of sociology. Perhaps the strongest yet tentative case that could be made based on this first analysis of abstracts is for the separation between quantitative/qualitative, science/humanities and macro/micro, where the former pole in each dichotomy seems to be more interconnected. Indeed, if we are to rely solely on what these word constellations suggest, one can even extend the question to ask if there might be a pattern in the form of the methodological

86 The reason no epistemological dichotomy is addressed is that it is the understanding of the analyst that practices are studied both from an objectivist and subjectivist point of view, as well as from a scientist and humanist one.
divide between quantitative and qualitative sociology found in the literature, but that one pole of this “master dichotomy” is more distinct. That is that words associated with one side of the divide – quantitative methods, macro level of society, and scientist vocabulary – are more frequently used together than words on the other side of the divide – qualitative methods, micro level of society, and humanist vocabulary – due to its dominant position in the largest network, and the fact that it has several smaller and dispersed word relations. To be able to assess this question, we will need computational text analysis methods specialized in determining if the corpus has any underlying semantic structure in terms of two or several distinct communities of words. This will allow us to establish more clearly whether there really is a kind of dual-paradigm represented by two schools in conflict over what sociology is and, if this is the case, how this relates to the methodological divide. Thus, the next step is to investigate if there are communities of words spanning across the abstracts with the help of cluster analysis.

Communities of Word Constellations

Continuing with the insights drawn from the previous section, this section also sets out to look into the most overarching word constellations of the abstract corpus by computing correlations represented as edges between nodes in a word network, but that add the feature of cluster analysis to further map out patterns formed by correlating words. The cluster analysis includes modularity, an algorithm that separates weak from strong correlations. By adding this feature, we are able to explore if there are distinct communities in the abstract corpus and to what extent they echo the two paradigmatic candidates of sociology, described in both the theoretical (e.g., Abbott, 2001) and the empirical (e.g., Schwemmer and Wieczorek, 2019) literature as being based on studies adhering to either a quantitative or qualitative methodology. In more technical terms, the technique is utilized to compute clusters of words where the most intercorrelated word constellations – i.e., word constellations with high modularity – form communities that can be visualized by the help of network analysis. Thus, with these computational techniques, one is able to draw attention to the most dominant

---

87 This analysis is performed by giving each word occurring in a sentence in these abstracts scores based on the words it co-occurs with. The most common words – such as between adverbs, conjunctions, and prepositions – are in most cases not meaningful for the analyst, just as was the case for the analysis presented in the previous section, which did not include stop words. Therefore, the analysis focuses on the most common nouns and noun phrases co-occurring in this abstract corpus.
Figure 3 is a visualization of the technique described in the previous paragraph. In this figure, we find that each word is coupled with a node, the size of which gives an indication of how often the word occurs in the corpus. For instance, we can see that the most frequently used words are here found in the tokens interview, data and concept. The edges connecting the nodes

---

88 Worth mentioning is that this exact method has been utilized on “all articles published in journals classified as sociology in the Web of Science in 2010-2015”, by Vincent Traag and Thomas Franssen in a blogpost published in 2016 under the title Revealing the quantitative-qualitative divide in sociology using bibliometric visualization (see link below). While many analyses of scientific publications that have used this method find multiple communities formed on the basis of research topics, Traag and Franssen found that their corpus seemed to follow another logic based on what appears to be a single divide running between two opposing communities interpreted as adhering to either quantitative or qualitative methodology. Interestingly enough, the authors also found that sociology journals were differently prone to publish articles resonating with either one of the poles (and not the other) or rather balanced between the two (https://www.cwts.nl/blog?article=n-q2v294 [received 2022-11-15]).
show how often words co-occur in the corpus, where thickness gives a simple illustration of how strong the correlations are. While the three mentioned tokens are all densely interconnected, a token like student at the very top of the graph seems, in contrast, rather isolated. Further, the nodes are colored in accordance with the communities that the cluster analysis found, which in this case are only two: orange nodes located on the left side of the graph and purple on the right side. Although both communities appear to be relatively interconnected, one might perhaps sense that the purple looks a little denser than the orange, which tentatively suggests that the words in that community occur together in a more systematic fashion.

When looking closer at the specific tokens constituting Figure 3, we find that many words are quite common to most social scientists – e.g., concept, framework, and extent – and therefore hard to draw any conclusions from. Indeed, the words in the middle of this figure that connect the two communities found in this corpus – e.g., gender, care, and century – are interpreted to be fairly ‘neutral’ in relation to the two communities (i.e., the quantitative and the qualitative community) since they occur relatively often together with words on both sides. Conversely, if we compare words located to the furthest left, for instance discourse (that is coupled with analysis) and construction, with words to the furthest right, say survey and sample, we start to get a hint of how the two different communities relate to certain sociological dichotomies. Indeed, in the following comparison of the communities, it will be suggested that most words coupled with orange and purple nodes, respectively, can be distinguished from each other on the basis of the poles that constitute one or a few of the dichotomies.

Based on the research questions posed at the beginning of this chapter, the most intuitive question to ask is what separates the orange nodes from the purple nodes constituting the figure or, more precisely, do the two communities resemble any of the poles adhering to a single or several dichotomies? Adding to the findings of the analysis presented in the previous section, the key candidate for a potential master dichotomy would from the offset be methodological, specifically that of quantitative/qualitative. Thus, we ask whether or not the word constellations forming the communities in Figure 3 are characterized by what theoreticians in the literature have suggested to be the dividing line between quantitative and qualitative sociology? Before we put this investigation into practice, it is imperative that we remind ourselves that any analysis of word occurrences such as this one cannot be used to expose distributions of methods or other related methodological indications for the sociology dissertations that this corpus of abstracts is based on. What we are able to see, however, is whether the two communities of words are often associated with either one of these crude characterizations of methodo-
logical approaches, as well as some of the other dichotomies they are thought to be associated with.

Based on the theoretical underpinnings explained earlier, the interpretation that will be put forward is that the communities detected by modularity and cluster analysis indeed appear to echo the poles found in some of the sociological dichotomies. When zooming in on the orange community on the left side of the visualization, we can, to begin with, locate the *quantitative/qualitative* dichotomy by a substantial set of words that are commonly associated with qualitative methods – *e.g.*, interview, interviewee, discourse analysis, and case study – and, in contrast, the purple community on the right side is populated with several words commonly associated with quantitative methods – *e.g.*, questionnaire, survey, respondent, and sample. A similar pattern appears to be traceable for the *macro/micro* dichotomy, since we can find that quite a few words commonly associated with the micro level of society – *e.g.*, interaction, practice, and *everyday life* – are coupled into orange nodes, while words alluding to the macro level – *e.g.*, country, welfare state, and labor market – can be found in the purple nodes. Further, regarding both the *science/humanities* and the *explanatory/interpretative* dichotomy, we find that the right, purple side of Figure 3 is populated with words associated with a scientific and an explanatory research model aimed at measuring causality and testing hypotheses – *e.g.*, effect, impact, cause, and hypothesis. In contrast, words related to a humanistic and interpretative research model, including study objects that are hard to quantify, occupy the left, orange side of Figure 3 – *e.g.*, interpretation, meaning, narrative, and feeling.

If the four dichotomies found in the figure are put together, we find on the left side the sociology dissertation abstracts that rely on words related to the *qualitative, micro, humanities,* and *interpretative* poles, while the right side of this figure is characterized by words commonly associated with the *quantitative, macro, science,* and *explanatory* poles. For the sake of simplification and for facilitating a possible interpretation, we can – like so many before us (e.g., Abbott, 2001; Erola et al., 2015; Kelle, 2005; Schwemmer and Wieczorek, 2020) – cling to a crude master dichotomy and from now on refer to the left side as the *qualitative community* and the right side as the *quantitative community*. Before continuing, though, it is of value to stress that five of the nine sociological dichotomies cannot as easily be mapped in the figure, including the three epistemological ones – *objectivism/subjectivism, fact/value,* and *theoretical/empirical* – and two of the ontological – *holism/individualism* and *structure/agency*. Thus, we seem to be
looking into a divide that can be understood primarily through the methodological dichotomies presented in the theory chapter.89

Based on the basic principle that word neighborship entails similarity in meaning, we could toy with the notion that the communal position of non-methodological words indicates the extent to which they are associated with qualitative or quantitative methodology. To put it in more technical terms, depending on whether the cluster analysis has assigned a token to the orange-colored community or the purple-colored community, we can make assumptions of the corresponding word’s association with the qualitative community and the quantitative community. From this position, we can also speculate that empirically-laden words appearing in the middle of Figure 3 – suggesting that the words are used equally often in both communities – are either “neutral” or “contested” in relation to methodology. For instance, at the top, we find the frequently used word gender and next to it words associated with common social institutions and problems – school, teacher, student, care, generation, equality, immigration – while the bottom of the figure includes words associated with time – century, decade, beginning. Continuing with word constellations found in the qualitative community, we find a set of orange nodes at the bottom of Figure 3 next to words that are interpreted to be associated with politics and movements – politic, actor, movement, regulation, authority, ideology – and next to it a set of words that seems related to organizations and professions – management, production, formation, staff, task, profession. Thus, we can suppose that these word constellations might be addressed more often in relation to words associated with qualitative as well as interpretative methodology, the micro level, and the humanities.

In comparison to the qualitative community, the quantitative community at the right side of Figure 3 appears to be constituted by more words with empirical connotations, which echoes the results presented in the previous section. At the top of the community we find, for example, words that seem to be associated with social identities and roles within family constellations – woman, man, family, mother, age, girl, child, parent. Continuing with the middle of the quantitative community, there are words that allude to comparisons of social differences such as social class and education – difference, class, division, outcome, education, status, risk, inequality, household, income. Moving to the bottom of the right-hand side, to the right there are words related to work – labor market, occupa-

89 The exception is here macro/micro that was labelled as an ontological dichotomy in the theory chapter but that has also been argued to be “a distinction between different levels of empirical reality” (Alexander et al., 1987, p. 1) and, thus, perhaps more methodological than ontological.
tion, employee, job, unemployment — and to the left words that can be linked with the welfare sector — welfare state, support, resource, service, access. Thus, we can this time assume that these word constellations are more related to quantitative as well as explanatory methodology, the macro level, and the sciences. Further, when taking into consideration this side’s pre-eminence in terms of correlated empirical words — regarding the number of words, the number of word occurrences, and the strength of word correlations — we also get an indication that the word constellations in the quantitative community seem to be comparatively more formalized than those in the qualitative community since many abstracts seem to refer to these specific word constellations and appear to do so with a relatively homogeneous vocabulary.  

In summary, these crude depictions of what the abstract corpus is characterized by as far as clusters of word correlations are concerned seem to suggest that the so-called “methodological divide” — argued by theoreticians and shown by other computational analysts to exist within the discipline — has left a mark on this corpus of abstracts for dissertations in sociology. Yet, it seems clear on the basis of this analysis that this particular divide is principally methodological and, thus, does not cover all theoretical divisions in sociology such as, for instance, epistemological dichotomies — objectivism/subjectivism, fact/value, and theoretical/empirical. To some extent, there appear to be constellations based on words with empirical connotations associated with each side of the divide, where, on the one side, the qualitative community seems perhaps more related to politics and organizations and, on the other side, the quantitative community looks to address indicators of social class and the welfare sector. Since the analysis presented in this chapter deals with the whole corpus, we will have to add more factors to this computation to evaluate how consistent this methodological divide is over time and place, which is needed to continue answering the research questions dealt with in this chapter.

---

90 Since this visualization of the words used in this abstract corpus is concerned with correlations between word counts, it is plausible that the abstracts associated with the quantitative community tend to use a more common vocabulary than those associated with the qualitative community. If we are allowed to speculate further, one potential reason why this computation shows a stronger association between non-methodological word constellations and the quantitative community could be that the sciences tend to deploy a more standardized terminology than the humanities. To illustrate, imagine that 40 dissertations study some aspect of crime, 20 of which are closer to the qualitative community and the other 20 to the quantitative community. If we say that the first half of the abstracts use a dispersed vocabulary and the other half a dense vocabulary, the visualization that the technique used here can generate would position the token crime much closer to words belonging to the quantitative community than the qualitative community.
Community Allocations over Time and Place\textsuperscript{91}

The interpretation of the cluster analysis of word correlations with network analysis presented in the section above suggested that the abstract corpus is separated into a \textit{quantitative community} and a \textit{qualitative community}, which is a synthesis of four dichotomies – \textit{quantitative/qualitative, science/humanities, explanatory/interpretative}, and \textit{macro/micro} – that might also be associated with certain empirical word constellations. In other words, it seems at this stage of the analysis, which has so far only dealt with the whole or the global level of the dataset, plausible to sustain studies in the literature proposing that sociology is divided into two methodological camps. To evaluate the sturdiness of this divide, we will have to analyze how it is expressed in different parts of the abstract corpus. Since theoreticians seem to put an emphasis on time and place as particularly important factors for how sociological knowledge changes (e.g., Abbott, 2001; Elias et al., 2002; Scott 2005; Turner 2001), these variables will guide this section. Thus, this corpus of abstracts for dissertations in sociology has been divided into smaller corpora on the basis of the year the dissertation was defended and the university where the defense took place. The first variable originally contains 40 different values – one for each year between 1980 and 2019 – but for the purpose of interpretability, the variable is aggregated into four comparably large groupings: 1980-1995 (n=198), 1996-2002 (n=199), 2003-2009 (n=200), 2010-2019 (n=218). The second variable is constituted by five values, one for each of the larger universities that since the mid-20\textsuperscript{th} century have been offering doctoral defenses in sociology.\textsuperscript{92} The five univer-

\textsuperscript{91} The variable of place has been added to the analyses that will be presented in the empirical chapters of this dissertation because place is believed to have an impact on the kind of sociology that sociologists conduct, both on a global scale and across different departments within a single country, as in this case (e.g., Aaltojärvi et al., 2008; Bjarman and Sifusdottir, 2002; Collyer, 2013a, 2013b, 2014; Moksony et al., 2014; Zougiris, 2019). From this it follows that one could expect that the place where one’s PhD has been conferred could have an influence in the kind of sociology that a PhD student would be inclined to conduct. Thus, although place is also presumably interesting for readers who are familiar with the national context, and would most likely be prone to draw conclusions on the different universities to which these analyses allude, it is important to state that this thesis not about Swedish PhD programs but rather about what the ‘product’ they generate (either in the form of an abstract or as full-text dissertations) to confer the degree suggests about what characterizes sociology. Thus, because place is hereby used as an angle for theorizing, rather than being primarily a backdrop to the results, information about the Swedish context is offered in Appendix A. Readers who are curious about the ‘peculiarities’ of Swedish sociology, the conferment of PhD degrees in this context, and the differences in university settings are welcome to consult the appendix in question.

\textsuperscript{92} As was stated in the method chapter of this dissertation and Appendix A, only abstracts for sociology dissertations defended in the five large universities are analyzed when the variable of space is studied. The reason for this is that the smaller universities have such a small number of sociology dissertations and most of them were first allowed to offer PhD programs in sociology a couple of years into the 21\textsuperscript{st} century. The small universities make up a total of 44 dissertations in comparison to the five larger universities’ 98-196 per university.
sities are, in descending order: *Lund University* (n=196), *Stockholm University* (n=186), *Uppsala University* (n=162), *Gothenburg University* (129), and *Umeå University* (n=98).

Thus, in this section, we will reuse the computation visualized in Figure 5, which served as the basis for the analysis put forth in the previous section, but add the variables of time and place to calculate what words are characteristic for each corpus and, in relation to the global level, what type of word constellations appear more frequently, less frequently, or around average. Returning to the clustered word correlation network analyzed above, the first task is to explore whether its division – interpreted as separating a *qualitative community* from a *quantitative community* – is consistent over time or if there are trend shifts. The analysis is presented by looking at how aggregates of dissertations derived from their defense year are distributed on top of the visualization presented in the previous section, Figure 3. This computation is presented in Figure 4, which is constituted of four graphs based on the clustered word correlation network – one for each time period (i.e., 1980-1995, 1992-2002, 2003-2009, and 2010-2019). In each graph, the nodes are colored on the basis of how often a word occurs in the corpus in that time period (i.e., the colors are not representing communities), and in the bottom of the figure there is a bar representing the spectrum of word occurrences in the abstract corpus, where green represents the average, yellow above average, and dark blue below average.

The first observation to make of Figure 4 is that the four graphs representing time share a similar overarching pattern: yellow, green, and blue nodes are spread out on both sides of the figure. This suggests that the separation between the *qualitative community* and the *quantitative community* presented in the interpretation of Figure 3 is not associated with a specific time period. Thus, to the extent that this method can capture the structure of the corpora, it appears that we can rule out the idea that one of the two methodological communities was more dominant in one period of time than in another (e.g., that 1980-1995 would be more associated with the *qualitative community* and 2010-2019 with the *quantitative community*). Rather, in terms of time, it seems that the so-called methodological divide is fairly consistent in the corpus since a pattern of frequently used words scattered on both sides of the graph appears, irrespective of which time period we focus on.

Taking a closer look at each of the graphs found in Figure 4, we see that certain words are overrepresented in different periods of time. This observation gives us an indication that some word constellations might be emerging and diminishing over time. Worth mentioning is that this method does not allow us to study the character, content, or manifestation of word constellations, which would require the use of a method like structural topic modeling.
introduced in the next chapter. Instead, what Figure 4 can show is only that the crude divide that the theories alluded to earlier, argued to be characteristic of sociological knowledge, seem to be leaving a mark on the words utilized in this abstract corpus irrespective of which time period we chose to zero in on. For instance, in the first time period (1980-1995), words for quantitative methods like survey and questionnaire seem to be used frequently, while their counterparts on the qualitative side, words like interview and discourse analysis occur less often in the abstracts analyzed. As for the second time period (1996-2002), words for both qualitative and quantitative methods appear more seldom. This pattern seems even more true in comparison to the third time period (2003-2009), where words of both categories seem to be used more frequently, suggesting a balance between the methodological communities alluded to earlier (and as far as abstracts are concerned). Finally, for the last time period (2010-2019), we can see almost the inverse pattern of the first time period (1980-1995): a frequent use of words such as interview and discourse analysis, but just an average use of the word survey and an infrequent use of the term questionnaire. Again, on the basis of these results, we cannot draw conclusions about how popular certain methods were at certain time periods. However, we can see how often these words interpreted to be associated with either the qualitative community or the quantitative community occur over time in the corpus.

Some initial remarks can also be made in relation to words with connotations to common sociological concepts by looking at their popularity in the abstract corpora of different time periods. For example, class and movement seem to be particularly popular in abstracts from 1980-1995, whereas the words labor market, occupation, and unemployment appear to spike in 1996-2002. Further, in 2003-2009, gender, man, and woman seem to stand out with high occurrences in the abstracts, and, in 2010-2019, practice, discourse, and inequality look to be at the forefront in terms of frequency. In contrast to these time-bound words, there are some that are consistently green – indicating an equal number of word occurrences throughout the corpora – like the three social institutions family, education, and care. As mentioned earlier, these are only tentative observations that would require the additional use of other tools to substantiate. Relying on abstracts also has its limitations, which is why we will be presenting analyses for the entire dissertation texts in the chapters that follow. Computational analyses of abstracts are, however, what is almost always done and, since they offer us an idea of the overarching word constellations that recur in how dissertations are presented to readers, they are the corpora for the first analysis that this dissertation relies on.
Figure 4. Temporal distribution of words in the word correlation network with cluster analysis and modularity (see Figure 3). The colors of the nodes indicate how frequently the words occur in each sub-corpus.
When moving over to examine the spatial aspect of the quantitative community and the qualitative community found in the clustered word correlation network, we will utilize the same procedure we used to generate Figure 4, with the only exception that the abstract corpus will be divided into corpora on the basis of space instead of time. As already noted, space is here measured in terms of the university where the dissertations in sociology that are being analyzed were defended, leaving us with five major candidates and five corpora of abstracts. The first thing we can notice when looking at Figure 5 is that the graphs do not follow the same overarching pattern as they did in Figure 4. Indeed, abstracts for sociology dissertations defended in the five universities in focus here differ from each other in terms of the distribution of word occurrences—i.e., the nodes are colored in different ways for each university graph. Nevertheless, a clear outline takes form when the abstracts stemming from the universities on the left side of Figure 5 are compared to those on the right side through the theoretical lens of the methodological divide.

On the left side of Figure 5, we find two graphs represented by the sociology dissertation abstracts that come from Uppsala University and Lund University. By looking at how the graphs of these universities are colored in the figure, we can sense that their dissertation abstracts include words that are commonly associated with both sides of the methodological divide. Yet, abstracts from Uppsala and Lund stand out from abstracts from the other universities by the fact that the nodes to the left are, to a great degree, colored in yellow-like colors and those to the right in blue-like colors. This suggests that sociology dissertation abstracts stemming from these two universities seem to have a leaning toward using words that could be classified as belonging to the qualitative community and not the quantitative community. Thus, perhaps the crudest interpretation of Figure 5 would be that dissertations defended at these two universities might tend to present a type of sociology that draws on the humanities, applies qualitative methods, studies phenomena on the micro level, and/or applies interpretivist methodology. There is, however, also a noteworthy difference between the abstracts from these two universities in terms of word occurrences. Abstracts for sociology dissertations defended in Lund seem to score particularly high for words of theoretical character (empirical/theoretical) that are associated more with the humanities than the sciences (science/humanities), like discourse, discourse analysis, concept, notion, and meaning. In comparison, the abstracts belonging to dissertations defended in Uppsala seem to be characterized more by words addressing primarily qualitative methods (quantitative/qualitative), such as interview and participant [observation], words with presumable connotations to the micro level (macro/micro) like everyday life, and, lastly, words echoing an interpretivist approach to research (explanatory/interpretative) like interpretation and narrative.
Figure 5. Spatial distribution of words in the word correlation network with cluster analysis and modularity (see Figure 3). The colors of the nodes indicate how frequently the words occur in each sub-corpus.
Turning to the right side of Figure 5, we find abstracts from two universities, with highly contrasting word occurrences to the universities on the left side, namely those accompanying sociology dissertations defended at Stockholm University and (to a slightly lesser extent) Umeå University. The abstracts belonging to dissertations defended at these universities appear to share the vocabulary of words found on the right side rather than the left side of the clustered word correlation network.\(^9^3\) Put differently, abstracts from these two universities seem to include a significant share of abstracts that use words strongly associated with the quantitative community rather than the qualitative community. Applying an analogous categorical interpretation, the words appearing in the dissertation abstracts stemming from Stockholm and Umeå allude to the (natural) sciences, quantitative methods, the macro level of society, and explanatory methodology. Yet, also in this case we find some notable differences between abstracts from these two universities. The abstracts for dissertations defended in Stockholm have a greater occurrence of words associated with the sciences (science/humanities) and research with an explanatory ambition (explanatory/interpretative), like hypothesis, effect, influence, sample, and distribution. Further, while both university corpora consist of abstracts that draw on words related to the macro level (macro/micro), Umeå’s abstracts seem to be associated more with welfare and social services – welfare state, service, equality, immigrant, and support – while Stockholm’s abstracts seem to be more about work and social inequality – country, labor market, occupation, inequality, unemployment, household, income, status, household, and child.

Following the reading of Figure 5 as illustrating a separation between overarching methodologies, abstracts from Gothenburg University appear to be at the crossroads of the qualitative community and the quantitative community. The intermediate position of these abstracts can be seen, for instance, in the fact that they score around average for both the words interview and survey as well as has higher than average occurrences for the words questionnaire and discourse analysis. In terms of more empirical word constellations, abstracts for sociology dissertations defended in Gothenburg have high scores for words alluding to professional fields like education and social services – profession, education, teacher, care, young person, and service – but also social inequality – labor market, division, and inequality – as well as words that resonate with taking social action – help, concern, task, regulation, and struggle – which might be oriented towards social values and policy (fact/value).

\(^{93}\) A detail worth mentioning that might explain why Umeå has such a low score on the token survey is that its abstracts seem to favour the word questionnaire.
On the basis of the crude analysis presented in this section, the supposed division between the *qualitative community* and the *quantitative community* seems more contingent on space than on time. It is true that the word scores varied, to some extent, at different time periods – perhaps some might even feel the urge to speculate on a tendency leaning towards the *quantitative community* over time – yet the interpretation presented here is that the two communities are rather balanced in terms of their distribution over the decades. In contrast, space appeared to be a much more important factor. In Andrew Abbott’s (2001) terms, one could therefore say the single sociology departments do not appear to be a microcosm of sociology at the global level – i.e., the university-specific corpora do not express the same pattern as the overarching national corpus of abstracts analyzed here. We found some abstracts from universities with patterns that resonate with the *qualitative community*, others with the *quantitative community*, and the abstracts of a single university seemingly placed right in the middle of the so-called methodological divide. This is an interesting finding in at least two senses. Seen from a local point of view, the abstracts from the dissertations defended at the five Swedish universities in focus seem to have different characteristics, which suggests that these sociology departments might differ in terms of their knowledge production. Further, from a global point of view, despite the supposedly different characteristics of the abstracts stemming from PhD dissertations in sociology defended at different universities, they do, in conjunction, look to reproduce the opposition seen in the literature on a national level. To sustain or reject the pattern found in the analysis presented in this section, we will now look into the content of each university’s corpora of abstracts and each time period’s corpora of abstracts.

**Word Constellations in Time and Place Corpora**

In this section, the abstract corpus is broken into smaller corpora based on the same categories as in the previous section, i.e., four categories of time and five categories of space based on the defense date of the dissertations and the defense place of the dissertations, respectively. Yet, this time we will reuse the same method that was applied on the global level in the opening section of this chapter, namely by computing separate word correlation networks based on each of the nine corpora. The reason for performing this technique is to unravel what word constellations are most general for every corpus so that we can investigate to what extent they reflect the categories’ respective positions in the division between the *qualitative community* and the *quantitative community* suggested above. Again, we will start by comparing the time periods, which are presented in Figure 6, and then move on to the universities, found in Figure 7. In both figures, the color of the tokens
indicates their occurrences in the corpus – starting with red (≈30), then orange (≈50), and ending in white (≈80) – and the width of the edges show the correlation between tokens – going from thin (≈0.3) to thick (≈0.7). Only words occurring in at least 15% of the full abstract corpus and that correlate at least 25% of the time are included in the visualizations presented in each figure.

Figure 6 comprises four graphs, each corresponding to a corpus consisting of dissertation abstract adhering to a certain time period: 1980-1995 (n=198), 1996-2002 (n=199), 2003-2009 (n=200), 2010-2019 (n=218). The graphs will be interpreted chronologically, starting with the 1980-1995 graph. While this graph, found in the upper-left part of Figure 6, appears to be strongly correlated, the abstracts within it do not seem to include many words adhering to any of the sociological dichotomies as presented in the theory chapter. The only exception is a few terms alluding to empirical studies (empirical/theoretical) and equally few that, perhaps, can be interpreted to have a small leaning towards the macro level (macro/micro). Indeed, to the left, we find a larger network structured by three interconnected tokens associated mainly with empirical research. First, we have the token results, leading to sweden, research, and importance that is linked to purpose. Second, the token data, which is extending only to empirical, and third, the token differences that is linked to swedish and class. We can here hint at a word constellation of social differences, which can also be found in two of the four dyads – power and influence, as well as economic and factors. The last two dyads seem to refer to more general sociological concepts that perhaps have a macro connotation – system and people, as well as form and time.
Figure 6. Word correlation networks computed on sub-corpora of the abstract corpus representing different time periods. The size of the nodes refers to how frequently words occur across the corpus (minimum 30 abstracts), whereas the thickness of the edges indicates how strongly two words are correlated (0.25-0.6).

In comparison to its predecessor, the 1996-2002 graph in Figure 6 consists of a larger network and smaller dyads and triads. In this case, we also find what appears to be an empirical focus (empirical/theoretical) as well as words resonating with both the qualitative community and the quantitative community mapped out in the previous sections of this chapter. Indeed, a good share
of the small networks is interpreted to refer either to qualitative methods
(quantitative/qualitative) – qualitative and interviews – or the micro
level (macro/micro) – life and situation, as well as experiences. Per-
haps we can here even sense an individualistic focus (holism/individualism)
in these abstracts with the tokens individual and individuals. In con-
trast, other smaller networks seem to pick up words alluding to the macro
level (macro/micro) when addressing politics – national, political, and
policy – as well as more general tokens that are seen in, for instance, the
dyad system and processes. Further, one small network is constituted by
words resonating with organizations – organization and organizational
– which are often conceptualized as meso-level phenomena. Similar to the
pattern found in the 1980-1995 graph, the largest network of the graph is
interpreted as rather ‘neutral’ in relation to methodology but includes many
words referring to empirical research (empirical/theoretical) – empirical
and material, as well as results and data – and what was previously
referred to as the rather “neutral triad” in the matter constituted by gender,
women, and family. Yet, by inferring the results of previous analyses we can
speculate that the abstracts in the largest network might have a minor leaning
towards a more macro level of analysis (macro/micro) as it includes the
tokens structure, level, and model, as well as labor and market.

The structure of the 2003-2009 graph shown in Figure 6 stands out from the
other three since it consists of more unique tokens and is made up of two big
networks, two middle-sized networks, and six smaller networks. The number
of relatively large networks suggests that we have more specific types of
word constellations that reappear throughout this corpus of abstracts. The
general interpretation to be presented is that the time period corpora of ab-
stracts have a balanced distribution of words adhering to the qualitative
community and the quantitative community, respectively. In the largest net-
work, situated at the bottom of the graph, we find the familiar triad gender,
women, and family, which seems to lie between words associated with qual-
itative methods (quantitative/qualitative) – qualitative and interviews –
and quantitative methods (quantitative/qualitative) – factors. The latter
token is further connected with differences and found, where the latter
joins a part of the network that seems to lean towards the dichotomies asso-
ciated with the quantitative community. This entails the tokens factors and
situation, leading on to words that are perhaps associated more with the
macro level than the micro level of society (macro/micro), namely people
and society, as well as labor and market, which in turn is connected to
data (empirical/theoretical). In turn, the token market is a link to the east-
ern part of the network, which picks up terms that in previous analyses have
been interpreted as referring to a macro-type sociology (macro/micro), with
welfare (presumably as in “welfare state”) leading to policy and public.
The token public is further linked with defined that is connected further to
concept, alluding to theory (empirical/theoretical), as well as individuals
and individual (that might perhaps be associated with individualism).

Further, the second largest network, positioned to the west, is interpreted as
consisting of one empirical side (empirical/theoretical) — starting with em-
pirical that is linked to understanding and material that is connected to
finally and structure (structure/action) — and one theoretical side (em-
pirical/theoretical) — with theories, leading to forms and form, as well as
theoretical and concepts, going over to field and central. Further, the
network of five tokens echoes a word constellation we recognize from the
other analyzes, the meso level, which is here centered around three intercon-
ected words with the stem organization — organizations, organization-
al, and organization — one of which leads to development and actors.
In contrast, the network consisting of a string of six tokens with related in
the middle appears to be split into words that are associated with the micro
level (macro/micro), namely experiences and experience, as well as a
scientific vocabulary (science/humanities), with influence, levels, and
level. Lastly, the smaller networks seem not to lean clearly towards any
side of the methodological divide but use general concepts for research, for
instance, research, results, and analyses, as well as main and ques-
tions. We also find a word constellation of life and conditions, and
another based on cultural, political, and discourse.

In the 2010-2019 graph of Figure 6, we can sense the common structure of
one big network and a few small networks of words deriving from the ab-
stracts, but which are sparser in this time period. As in the initial analysis of
the corpus on the global level, some of the small networks seem to be lean-
ing towards words associated with qualitative methods (quantita-
tive/qualitative) — qualitative and experiences — and the micro level of
society (macro/micro) — practice and practices, as well as everyday,
life, and society. There are also two small networks that sustain each side
of the empirical/theoretical dichotomy with, on the one hand, empirical
and material, and, on the other hand, theory, theoretical, and frame-
work. Again, we find the somewhat neutral triad — women, gender, and famil-
ly — which forms a triangle of its own. As for the large network, the right
side includes a constellation of words alluding to the sciences rather than the
humanities (science/humanities), made up of the tokens factors, im-
portance, related, and effects. These tokens are in turn connected to
words associated with empirical research (empirical/theoretical), e.g., re-
sults, linked to data and level. To the left, we can see a combination of
the tokens individuals, individual, and relationship, which is harder
to categorize but seems to be associated with individualism (hol-
lism/individualism), again, a set of tokens — examines, labor, and market—
referring to the labor market and thus presumably associated with the macro level (macro/micro).

Now, let us turn our attention to the spatial aspect of the abstract corpus, represented by five corpora based on the university where the dissertations were defended, namely: Gothenburg University (n=129), Lund University (n=196), Stockholm University (n=186), Umeå University (n=98), and Uppsala University (n=162). One of the first things to note when comparing the five graphs constituting Figure 7 is that, in comparison to the graphs based on time, they are far from equally structured. Rather, there seem to be three types of graphs: a few minor networks, one major network, or one medium-sized network as well as some smaller networks. This crude categorization will be the guiding principle for how the analysis in this section is presented.

Two of the graphs in Figure 7 stand out from the other three. These graphs, which share a similar structure of a handful of sparsely connected networks consisting of two to five tokens, adhere to the abstracts from the two universities interpreted in the previous analysis to be the ones using most words associated with the qualitative community, i.e., Lund University and Uppsala University. Since this method investigates word correlations, the results suggest that similar word constellations are seldom used in multiple abstracts. Yet, when comparing the relatively few tokens appearing in the two graphs and the networks they form, these networks still hint at the most dominant word constellations appearing in these abstracts. They include, for instance, words understood to refer to theoretical as well as empirical inquiry. For instance, empirical in the Uppsala graph only correlates with consists, but in the Lund graph, the same token (i.e., empirical) is connected to both material and analysis, as well as theoretical, which leads further to the token concept. In the Uppsala graph, we also find the token analysis that here is linked to main and focus, respectively. Another token of particular interest for its clear coupling with qualitative methods (quantitative/qualitative) is interviews that correlates with based in both university corpora, but in Uppsala’s case, it is part of a small network that also includes the tokens context and sweden. The abstracts stemming from Lund form a different graph. This one has a dyad constituted by understanding and life that perhaps could be argued to resonate more with an epistemology based on interpreting rather than explaining (explanatory/interpretative). There are further relatively neutral sociological concepts, seen in the tokens social and relations appearing in Uppsala’s graph, as well as the more thematically-specific token economic that is coupled with development and political, respectively.
Figure 7. Word correlation networks computed on sub-corpora of the representing universities. The size of the nodes refers to how frequently the words occur in the corpus (minimum 20-39 abstracts), whereas the thickness of the edges connecting them indicates how strongly the words are correlated (0.25-0.6).
In contrast to the two graphs addressed thus far, those of Stockholm University and Umeå University are both constituted by only a single large network (except an outlying dyad in Stockholm’s case) consisting of around 30 tokens. This suggests, in stark contrast to Lund and Uppsala, that a substantial share of the abstracts in these corpora use a similar vocabulary. This can be seen in, for instance, the high occurrence of empirically-laden words found in tokens like empirical and data, and the absence of any words referring to theory (empirical/theoretical). In the previous section, it was suggested that abstracts stemming from both universities are associated more strongly with the quantitative community than the qualitative community. Yet, looking at the graphs in Figure 7, only the abstracts constituting the Stockholm graph include a word directly referring to quantitative methods, i.e., survey. Perhaps we might allow ourselves to infer from tendencies found in previous analyses that the token data – which is found in both graphs and correlates with survey in the case of Stockholm – is leaning slightly towards the quantitative pole (quantitative/qualitative). We also find the token factors in both graphs and, in the case of abstracts coming from Stockholm, the tokens effects and influence; these three words can be argued to also be associated with quantitative methodology, as well as to echo a scientist rather than humanist vocabulary (science/humanities). Abstracts for dissertations defended in Umeå are instead characterized by an exception to this pattern in the tokens interviews and context that allude to qualitative methodology (quantitative/qualitative). While these tokens have a peripheral position in the network, they seem to sustain the previous interpretation that abstracts stemming from Stockholm might be associated even more with the quantitative community than those for dissertations in sociology defended in Umeå.

Further, we can find in the graphs of both Stockholm and Umeå word constellations associated with certain research topics that were also found in the previous analyses, such as the economy – seen in the token economic that is linked to class, conditions, and resources, as well as development and conditions – and work – the interconnected tokens labor and market (as in “labor market”). In the case of abstracts coming from Stockholm, the work-related tokens also correlate with effects, women, gender, and family. Since this triad is further connected with the tokens countries and difference, and if labor markets are conceived to be macro phenomena, one could perhaps propose the interpretation that this word constellation alludes to the macro level of society (macro/micro). Looking at the densely connected southern part of the network constituted by abstracts stemming from Umeå, the evidence suggests that a focus on the macro level is even more plausible but this time it seems coupled with politics rather than work, which is seen in the linked tokens sweden, national, level, welfare, support, public, policy, and political.
Lastly, in the case of abstracts from Gothenburg University, the graph again seems to be structured in a way that lies somewhere between the two graph types generated through the abstracts originating from dissertations defended in Lund and Uppsala, on the one hand, and Stockholm and Umeå, on the other. Indeed, this graph of word correlations computed on abstracts from the Gothenburg corpus comprises both a relatively substantial but sparse network of 17 tokens and several smaller networks of two to five tokens. From the offset, this seems to echo the interpretation of abstracts from dissertations defended in Gothenburg that was put forward in the last section, where it was suggested that they seemed to be positioned somewhere in the middle of the qualitative community and the quantitative community. This understanding seems to make sense when investigating the graph’s word constellations that the abstract corpora generate. For instance, we find that the larger network is centered around the token main that is correlating with the tokens central, economic, and empirical (empirical/theoretical), where the latter token is correlating with relation that leads to theoretical, position, and strategies (empirical/theoretical). Further, the token economic is leading to life, time, terms, and interviews (quantitative/qualitative), whereas the token central is connected to related, theory, political, gender, and women (empirical/theoretical). In addition, two small networks are interpreted to be associated with empirical studies, one starting from the central token results that leads to data, analyses, and knowledge, and the other, constituted by understanding and studies. Further, there are small networks that might be interpreted as partly being slightly more associated with the macro level than the micro level (macro/micro) – one constituted by labor, market, and public, leading to means and change, the other by the token cultural that is linked to society, processes and aim, which is correlating with based. Seemingly supporting the understanding of sociology dissertation abstracts from Gothenburg as taking a middle position, there is also a small network based on words that allude to the meso level made up of the tokens organization, development, and education. Finally, formed by commonly used concepts in research, which are somewhat neutral concerning the sociological dichotomies, are focus, purpose, social, and role.

To summarize the findings presented in this section, the overarching interpretation put forward is that the word constellations with the highest prevalence in each abstract corpus reflect the patterns found in the previous analyses of correlations between words, suggesting that the overarching separation between a qualitative community and a quantitative community seems fairly stable over time, but more contingent on space. Beginning with the analysis of corpora divided into time periods based on Figure 6, the word constellations seem to echo those on the global level visualized in Figure 2 that was presented at the beginning of this chapter. In general, the basic ar-
rangement includes a larger and denser network of words associated with the *quantitative community*, which sometimes include a small contribution alluding to the *qualitative community*, as well as several smaller networks constituted by words resonating with the latter community. Thus, the structure and content of the word constellation found in this analysis seem to sustain the previous interpretation that the so-called methodological divide does not appear to be that contingent on time, and adds more substance to the speculation that the *quantitative community* might be more coherent than the *qualitative community* in terms of using a more standardized vocabulary. Seemingly, the only exception to these general tendencies is the graph 1980-1995 in Figure 6, which shows surprisingly few correlated words and even fewer associated with any of the sociological dichotomies.

In vast contrast, the word correlation networks generated on the university corpora of abstracts, which is visualized in Figure 7, all appear to differ from the global corpus. While abstracts for dissertations defended in universities like *Stockholm* and *Umeå* show graphs that consist of one large, interconnected network, the graphs generated by abstracts originating from *Uppsala* and *Lund* comprise a few smaller networks. Further, the abstracts from the two former universities are interpreted as drawing more on words associated with the *quantitative community*, while some of the few words appearing in the graphs of the two latter universities shared connotations with the *qualitative community*. Taking into consideration the structure of the graphs associated with the respective communities, this analysis seems to support the speculation raised earlier in this chapter, i.e., that the *quantitative community* is associated with a more coherent vocabulary than the *qualitative community*. The reasons why words associated with the *quantitative*, *explanatory*, *science*, or *macro* poles tend to use fewer synonyms, more often address similar phenomena, or use the same concepts, is debatable, and would need other computational text analysis tools to be investigated. Nevertheless, at this stage of the analysis, it seems plausible that space plays a role in dividing the corpus into two different methodological camps, which seem to be rather balanced on the global level, and relatively stable over time. The analysis presented in this chapter has also hinted that abstracts for sociology dissertations from different time periods and universities might be more strongly associated with different themes, which will be studied in more depth in the next chapter with the help of topic modeling.

**Closing Thoughts on Word Constellations**

As was stated at the beginning of this chapter, its main focus has been on investigating three research questions, one on the so-called fragmentation of
sociological knowledge, another on unravelling whether there are paradigms in sociology, and a third on the social conditioning of sociological knowledge. The strategy for tentatively investigating the first research question in this chapter was to generate word correlation networks on the abstract corpus in search of recurring ‘word constellations’ that might indicate that sociological knowledge is or is not fragmented. The second research question was tentatively investigated by searching for overarching word constellations that might articulate paradigms. Following the reasoning presented in the theory chapter, paradigms are operationalized as repeatably referring to a set of poles in the recurring dichotomies of sociology. Further, the social conditioning of knowledge was scrutinized by looking into how these overarching word constellations are contingent on time and place. This closing section will begin by taking a more general view of the results of what the analyses conducted managed to illustrate, and successively move into suggesting a tentative answer to the third research question, and considering what this answer might tell us about the crisis of sociology.

In the analyses performed on the global level of the corpus, it was suggested that while words with connotations to phenomena like welfare states and labor markets were recurring, the overarching word constellations seem to address different types of research. When adding a clustering algorithm with modularity, the word correlation network was divided into two clear parts. The principal word constellations of each part seemed to be persistent in the timeframe from 1996 to 2019 but less so between 1980 and 1995. Yet, when adding the periods together, the word constellations, as well as their division, were found to be relatively durable over time. Thus, it appears that the content of the abstract corpus is most diverse in the early period but still relatively integrated throughout 1980-2019, particularly so for the later period, which suggests that sociological knowledge is not becoming increasingly fragmented.

In terms of place, however, the abstracts stemming from different departments of sociology did not seem to reflect the pattern found for the whole corpus. Instead, abstracts for sociology dissertations defended at two universities, *Stockholm* and *Umeå*, had a dense vocabulary, represented mainly by a single interconnected network and primarily picked up words associated

---

94 Research question one reads: If fragmentation is understood as a historical process where the content of the knowledge produced in a discipline becomes increasingly disintegrated, do the corpora offer evidence that sociological knowledge is in a state of fragmentation?

95 Research question two reads: Given that a paradigm denotes a crude replica constituted by reiterations of the same crude analogies in a discipline, do the corpora offer evidence that there are one, two, multiple, or no paradigms in sociology?

96 Research question three reads: Following the notion that the social conditioning of knowledge is expressed in the time and place of its production, do the corpora offer evidence that sociological knowledge is socially conditioned?
with one side of the divide. In contrast, abstracts stemming from two other universities, Uppsala and Lund, encompassed words associated with the other side, yet only a few of these words were shared between the universities. The abstracts connected to the last university included in these analyses, Gothenburg, were found to be right in the middle of the divide, with a corpus showing both larger networks of correlated words and a few dispersed tokens alluding to both sides of the divide. Based on this analysis, sociological knowledge appears to come in patterns that are recurring over time but conditioned by the place in which it is created.

The sociological dichotomies presented in the theory chapter were found to provide a good set of tools for sorting out the word correlation networks and mapping out patterns of sociological knowledge. Indeed, it was suggested that the methodological dichotomies served as the best candidates for conceptualizing the seemingly robust division seen between word constellations of different sorts. Thus, the main line of reasoning proposed by this chapter is that the divide ran along four principal dichotomies – science/humanities, explanatory/interpretative, quantitative/qualitative, and micro/macro – that together formed two communities. Based on the idea of a ‘methodological divide’ serving as the focus in several theoretical works (e.g., Abbott, 2001; Azarian, 2022; Turner, 2015), the communities are labelled the quantitative community and the qualitative community, where each comprises a ‘master dichotomy’ synthesizing the poles of the four dichotomies. Similar patterns have been found in computational text analyses of journal abstracts in international sociology journals (Schwemmer and Wieczorek, 2020) as well as hand-coding of sociology journals in the Nordic countries (Erola et al., 2014) and Great Britain (Payne et al., 2004).

In other words, the results presented here, which are based on analyses of abstracts for dissertations in sociology defended in Swedish universities, resonate with previous research suggesting that sociology is characterized by a so-called methodological divide. This methodological divide appeared to be fairly sturdy in terms of time, but much more contingent on place. Indeed, these analyses suggest that abstracts stemming from different sociology departments draw on different word constellations. Two universities, Stockholm and Umeå, were interpreted to be the ones scoring the highest for the quantitative community. These universities further appeared to have the most even vocabulary. In comparison, the universities suggested to be most associated with the qualitative community, Lund and Uppsala, had the least correlated words.

In the frame of the operationalization of paradigm theory guiding this dissertation, the methodological divide can be reimagined. Sociological knowledge, as it is expressed in the abstract corpus, can be understood to
reflect a *dual-paradigm science*. The intuition behind this interpretation is that the material suggests a recurrence of sociological dichotomies adhering to *crude replicas* of two separate paradigms. Thus, at this stage of the study, we find ourselves with the tentative picture of sociology as a *dual-paradigm science*, governed by a methodological divide that is contingent on place. Perhaps this divided aspect of sociological knowledge (seemingly based on methodological preferences) might be related to the experienced crisis of sociology. Yet, three more analytical chapters remain in which to unravel whether the patterns suggested in this chapter hold their ground. Thus, the analyses presented in the next chapter will target the abstract corpus from another angle by conducting a computational technique called structural topic modeling.
7 Thematic Arrangements

This chapter is a continuation of the tentative investigation set out in the previous chapter of the three research questions, which investigate fragmentation, paradigmatic status, and social conditioning of sociological knowledge. In this chapter, these questions are tackled by exploring the consistency of ‘thematic arrangements’, which is conceptualized as combinations of the most prevalent topics in the corpus. The corpus of study is again the 815 sociology abstracts for dissertations in sociology defended between 1980 and 2019 at Swedish universities that confer the PhD degree in sociology.

One major difference between this chapter and the previous chapter on word constellations is that a human could, in principle, calculate the word correlations by hand while the topics can only be divulged by applying a set of unsupervised machine learning algorithms, often referred to as topic modeling. Within computational text analysis, topic modeling is a task performed to discover the topics that the machine estimates to be the best description of a set of documents. The method is about leveraging computational power to generate latent topics (only existing within the data itself) from a large corpus in a matter of minutes. One important feature in which topic modeling differs from, for instance, hand-coding generated themes is that the former can be performed “with a minimum of human intervention”, and this, some researchers have argued (Mohr and Bogdanov, 2013, p. 546), “makes the method more inductive than traditional approaches to text analysis in the social and human sciences”.

Presumably, the best documented and most applied forms of topic modeling are based on Latent Dirichlet Allocation or LDA (Blei et al., 2003). In simplified terms, LDA generates a fixed number of topics from the corpus in

---

97 Research question one reads: If fragmentation is understood as a historical process where the content of the knowledge produced in a discipline becomes increasingly disintegrated, do the corpora offer evidence that sociological knowledge is in a state of fragmentation?

98 Research question two reads: Given that a paradigm denotes a crude replica constituted by reiterations of the same crude analogies in a discipline, do the corpora offer evidence that there are one, two, multiple, or no paradigms in sociology?

99 Research question three reads: Following the notion that the social conditioning of knowledge is expressed in the time and place of its production, do the corpora offer evidence that sociological knowledge is socially conditioned?
which its words are sorted based on their probability of occurring in the topics, as well as the probabilities of the topics occurring in each document. On this basis, every document can be described as a distribution of topics sorted by their probability of occurring in the text. This analysis utilizes the Structural Topic Model (STM) (Roberts et al., 2013), which builds on the classic LDA model but adds the potential of computing the effect of covariates or metadata on the topics. This combination, the authors argue, is particularly valuable for the social sciences. After a set of statistical methods and an intensive exercise in human interpretation, 20 topics were deemed to be the perfect fit to cover the semantical landscape of the corpus. See Appendix C for a more technical description of topic modeling, how the STM utilized in this chapter was trained, and how the data were pre-processed.

To investigate fragmentation, the focus is on how the topics are distributed over time, particularly whether there are fewer topics in the 20th century and more in the 21st century, or few recurring topics (indicating fragmentation), or perhaps a ‘core’ of topics that remain intact over time (implying no fragmentation). Paradigms are investigated through the theoretical lens applied in this dissertation, namely the sociological dichotomies presented in the theory chapter. While all nine dichotomies are applied to unravel the thematic formations emerging from the analysis, the results of the previous chapter have moved those associated with the ‘methodological divide’ (separating a quantitative community and a qualitative community) to the forefront. Further, the metadata explored are once again time, the year the dissertations were defended, and place, the departments of sociology (coded as universities) where the abstracts for these dissertations were defended.

The chapter is based on a set of visualizations derived from a single STM based on 20 topics computed on the complete corpus. The presentation of this analysis is divided into four sections. The first gives a general overview of the topics by interpreting the words they are constituted by and their prevalence in the abstract corpus, and the second explores how the topics relate to each other through a network analysis of correlations. In the third and fourth sections, the effect that time and place have on how the topics are distributed in this abstract corpus is investigated. At the end of the chapter follows a quick discussion of how the results generated from the analysis can help answer the three research questions presented at the beginning of this chapter.
Interpreting and Labeling Topics

To reach an overarching understanding of the primary thematic arrangements in the corpus, this section sets out to describe the 20 topics modeled on the abstract corpus by the structural topic model. In practice, this task is centered around labeling each topic based on its most apparent tokens. Again, using this type of method, each topic can be described by a distribution of words (spanning from the most probable to the least probable words to occur), whereas the documents, which in this case are abstracts (n=815), can be described by a distribution of topics (n=20). The first statement, the probability that words will occur in topics, is encapsulated in the model’s beta value (β), and the second, the probability that topics will occur in documents, in its theta value (θ). This section will deal primarily with trying to understand the latent topics by assessing their most probable and distinct words, yet these values are viewed in the light of each topic’s overall prevalence in the complete corpus. To interpret and label the topics, three measurements are applied. The primary one is β, capturing the most probable words for the topics, which is visualized together with the topics’ prevalence in the only figure of the section, Figure 8. The beta is supplemented by the FREX, which is a weighted harmonic mean of the word’s rank within a topic concerning its frequency and exclusivity, and Lift, which highlights topic-specific words that occur less frequently in the corpus by providing word weights that are the sum of a division between their frequency to occur in a specific topic vis-à-vis the other topics (cf. Bischof and Airoldi, 2012). Again, to provide an answer to the research questions guiding this chapter, all topics will be interpreted through the theoretical lens of the sociological dichotomies.

In Figure 8, we find a visualization of the 20 topics, where each topic is assigned a bar of a unique color. The length of the bar is based on the topic’s prevalence in the corpus, that is the documents’ theta values (θ). The most general topic occurring in all the abstracts is the one named Theory/Knowledge and the least is Problems/Treatment. These topic labels have been reached after a thorough investigation of all words associated with every topic, which is described in detail in this section. To get a hint of what words have the highest β value in all topics, there is a short list in the figure of the most probable tokens to appear in a topic next to its colored bar. The tokens are stemmed versions of the original words occurring in the abstract corpus, meaning that only the stem of each word is kept, which is done to fit similar words into the same token and give a better representation of the most general thematic arrangements in the corpus. On the basis of Figure 8, the process of labeling the topics will be described one topic at a time, be-
ginning with the topic that has the highest $\theta$ value, *Theory/Knowledge*, and ending with the lowest $\theta$ value, *Problems/Treatment*.

Figure 8. The 20 topics of the abstract corpus generated by the structural topic model, presented with their assigned labels and a small share of their most contributing words. The gamma refers to the topics’ prevalence across all documents, i.e., their percentage share of the corpus (where the 20 topics total 100%).

Starting at the top of Figure 8, we find a topic named *Theory/Knowledge*. Again, the position of *Theory/Knowledge* to the other topics indicates that this is the most prevalent topic in the entire abstract corpus. From the words to its left—concept, knowledge, science, theory, culture—which are the most probable to occur in this topic, we get a first sense of what kind of meaning this topic constitutes. The word *science* is also present in the FREX and the Score, but here we also find words associated with philosophical concepts, as well as names of well-known theoreticians. For the former, we have ontolog, phenomena, thought, weber, and methodolog, and, for the latter, adorno, philosophi, freud, weber, and schutz. This pattern is sustained by the Lift words—freud, kant, relativist, creatur, human, metaphys, presupposit. Thus, the proposed reason for naming the topic *Theory/Knowledge* is based on the interpretation that these abstracts, on the one hand, cover instances of words related to the theoretical domain—theories, concepts, phenomena, thought, etc.—which is here linked to the *theoretical* pole of the *theoretical/empirical* dichotomy presented in the theory chapter. Further, most of the names raised in the FREX and the Lift, which
pick up less probable but more specific words for the topic, are interpreted to relate to theoreticians that have shown a certain interest in questions of human entities when theorizing on societal entities – Max Weber, Sigmund Freud, and Alfred Schütz100 – and, thus, seem to lean toward the second pole of the ontological dichotomies (*holism/individualism*, *structure/agency*, and *macro/micro*). On the other, the topic seems to include specific references to forms of knowledge (philosophy, science, culture, etc.), as well as philosophical branches (metaphysics), authors (Immanuel Kant), and positions (relativism), which added together seem associated more with the humanities than the sciences (*science/humanities*). Nevertheless, the high prevalence of the topic *Theory/Knowledge* in the corpus is interpreted as the common reference to sociological theory in the abstracts, rather than a specific branch of theoretical sociology.

Below *Theory/Knowledge*, we find the second most prevalent topic, *Gender*. Looking at the words most likely to appear in this topic, there seem to be two categories: generic words for gender types – *women*, *gender*, *men*, *female*, *male* – and classically sociological as well as more feminist concepts – *interview*, *sexual*, *equal*, *experi*, *construct*, *power*, *feminist* – that seem to be coupled with qualitative research methods (*quantitative/qualitative*), as well as an epistemology focused on human subjects (*objectivism/subjectivism*). Almost the exact same words appear in the FREX and the Score, with the addition of *masculin* and *motherhood*, while the Lift brings about words related to sexuality that are interpreted to fall within a feminist discourse – *gaze*, *gender-neutr*, *heteronorm*, *homosexu*, *jargon*. Following this interpretation, the topic is named *Gender* to cover both references to gender types (*women*, *men*, *males*, *females*, etc.) that are not clearly associated with any dichotomy, which explains the topic’s high prevalence, and a specific theme conceived to be associated to feminist theory and the field of gender studies, which might resonate more with the *qualitative community* than the *quantitative community* found in the previous chapter.

Moving on to the third most prevalent topic in the abstract corpus, *History/Economy*, we find that the words with the highest probability appear to have historical and economic connotations – *system*, *modern*, *econom*, *period*, *histor*, *industr*. The words suggested by the FREX and the Lift scores sustain this reading as well, and seem to add a common historical sociology theme surrounding the concept of modernity following the shift from agriculture to industrialization with Karl Marx’s theories at the forefront, with words like *industry*, *modern*, *compani*, *marxism*, *marx*, and

100 The German letter Ü is replaced with U in the abstract corpus since some authors have chosen to use the Anglified form of German words and others have not.
agriculture. While economy is hard to scrutinize using the prevailing dichotomies of sociology, there is a high frequency of words that can be interpreted as related to history—modernity, period, history, industrialization, and agriculture—and coupled with Marxism they would seem to be more related to the humanities than the sciences (science/humanities). Nonetheless, considering the thematical breadth and the societal embeddedness of the words characterizing this topic, spanning from history to economy, it would make sense that History/Economy appears so commonly in this corpus of sociology dissertation abstracts.

The top tokens of the fourth most prevalent topic Welfare/Countries—policy, state, welfare, polit, institut, parti—seem to be constituted foremost by words to the political sphere (policies, politics, parties, etc.) and the welfare state. The notion of welfare appears to hold for the remaining words with the highest probability—support, public, benefit—but here the topic also picks up the words attitude, country and European, which together are interpreted to indicate a focus on the macro level (macro/micro). One speculation would be that these words refer to some form of comparative perspective across different (presumably European) countries on the topics of states, policies, and welfare. This reading is supported by the FLEX words and Score words that, on the one hand, pick up state, policy, and welfare, and, on the other, the words cross-national, country, European, and regime. In either case, it looks like the Welfare/Countries topic comprises words relate specifically to welfare, and countries in general, and it is within the framework of macro that the topic’s words with political connotations are construed.101

The succeeding topic found in the sociology dissertation abstracts, Organization/Movement, is dominated by one chief word, organ, which in this corpus primarily forms words with the stems “organization” and “organize”.102 This word is coupled with the similar word organiz and, looking through the FLEX and the Score, we find the same two words as well as two more acquaintances to organizational matters: bureaucraci and bureaucrat. The second most contributing word is movement, which is interpreted as different forms of social movements, which also tend to be organized. Thereafter we find a set of words in the beta, the FLEX, and the Score that are not as probable to occur and that refer either to political matters—polit, member, democrat, activist—or have a theoretical leaning—structure, model, theory, concept. Both organizations and social movements can be con-

101 As will be shown later in this section, there is another topic called Discourse/Politics that is also centred on the token polit but that is interpreted to pick up the more micro aspects of politics.
102 The token organ appears only one time in a context referring to the human body and another time to a social institution.
strued as meso level phenomena in the middle of the macro/micro dichotomy. Perhaps we can, with the help of the Score, also sense a theoretical vocabulary (empirical/theoretical) and references to social structures (structure/agency), but this interpretation is made with great caution and should only be read as in parenthesis. The full-text analyses that will be conducted in the next two chapters may be able to offer us some indication of the validity of this interpretation.

The most probable words for the next topic, Discourse/Politics, are discourse and polit, followed by some words that intuitively seem related to both words – debat, conflict, discurs, public, analys, issu. Further on, we also find fewer words like ident and practic that could be conceptually coupled to discourse in the sense of discursive identity and practice, but which are also broader concepts used outside of these two realms. The same goes for nation and refuge, which could be interpreted as popular discursive and political matters in the corpus, yet, of course, are not exclusive to just these aspects of the phenomena. The topic of Discourse/Politics as generated from the abstracts is understood as related to a more interpretative type of methodology like discourse analysis (explanatory/interpretative) that, in contrast to the seemingly macro-laden Welfare/Countries topic, picks up more micro-level aspects of politics (macro/micro).

Thereafter, we find a topic labelled Labor/Market, which appears based on words that can be sorted into two strongly overlapping categories. The first is words with connotations to labor and work – labor, unemploy, employ, employe, worker, job, work, workplace – and the second to the economy, particularly the notion of markets – market, wage, econom. As we saw in the previous analysis, the concept of the labor market – encapsulating all societal forms of work and employment – is a prevailing theme in itself within this abstract corpus, and is understood to refer to the macro level (macro/micro). But other types of markets based on economic relations that are not necessarily macro phenomena seem to occur in the topic as well, as can be seen from the words with the highest Lift scores – meat, precar, shrink, textil, livelihood, recess, submarket. Thus, the label Labor/Market is thought to encapsulate multiple references to work, as well as a wider arrangement of economic themes beyond the labor market.

Service/Profession is another topic found in the topic model and is constituted by two main themes. One is encapsulating highly probable words related to social services, particularly care and medicine – care, medic, servic, plan – and the other, professions including management – manag, profession, strategi. This pattern is supported by the words with the highest FREX, with one set that can be related to similar forms of social service –
nurs, medic, patient, municip, care — and the other relating to matters of professions — profession, staff. None of the words associated with the topic seem to be directly or indirectly associated with any of the sociological dichotomies informing the interpretation of results generated through these computational text analysis methods.

Zooming in on the next topic, Qualitative, its words with high β scores — particip, situat, interview, collect, environ, condit — are interpreted to share connotations with empirical research (empirical/theoretical), qualitative methods like interviews and participant observation (quantitative/qualitative), and what is interpreted to be a focus on the micro level (macro/micro) based on words like conditions, situation, and environment. Further, there are more specific words — activ, union, attitud, power — that are harder to bind with a single thread but that carry connotations to power, attitudes, activism, and unions. Based on this interpretation, which sees unity in the concepts addressing qualitative research and diversity in the remaining words, the topic is labelled Qualitative. Regarding the so-called methodological divide mapped out in the previous chapter, this topic appears to be situated rather clearly in the qualitative community.

The topic of Micro/Humanist is interpreted to be constituted by words that seem to be centered around humans in social research, from their actions to their thoughts. First, we have words that seem related to the micro or interactionist level of society (macro/micro) — interact, practic — and second, general words for theoretical knowledge (empirical/theoretical) that is interpreted to be associated with the humanities (science/humanities) — understand, theori, concept — including emergent or relational concepts like process and network. The other three algorithms which have been used in these analyses — i.e., the FREX, the Lift, and the Score — are all understood to suggest words associated with an individualist perspective (holism/individualism), like concepts adhering to social psychology, whether in its classical form associated with George Herbert Mead — mead, self, gestur, interact, play, game, communic, self-reflection — and concepts accompanying more modern references in micro sociology (macro/micro) like Thomas J. Scheff and Judith Butler — scheff, butler, bodili, boundari, undecid. Since it seems that the words resonating with the micro level are here accompanied by humanistic rather than scientific methodology, the topic is given the name Micro/Humanist.103

Turning to the topic Family/Background generated by the STM modeled on the abstract corpus hereby analyzed, we find that the β levels suggest that

103 An example of a more scientific or positivistic approach to the micro level would, perhaps, be ethnomethodology (cf. Tilley 1980).
families is by far the most probable word, and at half that token’s score are children and parent, which together with similar words – child, father, women – are interpreted relating to family relations. These tokens all appear in the FREX and the Score, in addition to the related words sibl and divorc. The fourth most probable word in this topic, which is occup, suggests that words like occupations are mentioned in these abstracts together with aspects of the family. This combination of words is interpreted as adhering to different markers of social backgrounds. This interpretation is supported by the Lift tokens since most of them seem to be in a similar ballcourt of social background indicators – carer, earner, machin, modest, poverty. Thus, the label Family/Background is a shortening of the theme of family and social background in this abstract corpus. It is, at this stage of the analysis, speculated whether this topic picks up words for large comparisons of populations and social structures (structure/agency) that would resonate with quantitative empirical research (empirical/theoretical and quantitative/qualitative) on the macro level (macro/micro), but this will have to be investigated further with other techniques.

The topic of Positivist/Risk is also interpreted as being divided into two parts. First, we find that the lion’s share of the most probable tokens seems related to scientific research (science/humanities) aimed at explaining, measuring, modeling, or evaluating the influence, effect, or impact of something – influenc, model, effect, indic, evalu, impact. These words can all be related to a more natural scientific or, if you wish, positivist vocabulary surrounding the idea that causality within the social world can be measured and explained (explanatory/interpretative), which is often (but not always) drawing upon a quantitative methodology (quantitative/qualitative). Second, the highly probable word risk suggests that perhaps there is a tendency to research risks, and looking at the other three algorithms – FREX, Lift and Score – we get an idea of what these kinds of risks could be – sick, retir, fertil. If positioned in relation to the methodological divide, which, as already mentioned, is guiding this interpretation, Positivist/Risk is a topic that would clearly fall within the quantitative community.

Looking at the most probable words for the topic of School/Media, it seems possible to categorize them into another thematic pair. There are some tokens with clear connotations to schools, like school and teacher, and some tokens interpreted to be associated with media and debatable subjects – media, moral, drug, polic. The same pattern appears in the other algorithms, where we find, on the one hand, the tokens curriculum, teach, pupil and peer and, on the other, the tokens news, oppon, speaker, and televis. The former set of words sustains that school is a prevalent theme in the topic and the latter suggests that media is another common theme. The speculation put forward is that words referring to morals, policing, drugs,
and so on are used in the same abstracts as those referring to schools or media. Nevertheless, the only dichotomy that seems to be related to the topic is that of fact/value since it picks up values. However, it is not clear if values here are studied from a value-oriented or a fact-oriented position, so it is suggested that the theme is not associated with any dichotomy.

*Space/Culture* is an example of a less distinct topic, which is why this is interpreted as not being directly related to the nine sociological dichotomies presented in the theory chapter. Some of the most probable tokens to occur in the topic seem to have connotations to more general aspects of space – environment, emerg, area, space – and community (including cities or citizens) – citi, communiti – as well as case and action. Further, we find more specific words with connotations to crime – crime, crimin – and an urban context with urban. By reading the dissertation abstracts through the lens of these themes, the topic is interpreted as capturing documents referring to various aspects of space. Looking through the FREQ, the Lift, and the Score, we find the same set of tokens but also words that are understood to encapsulate different aspects of culture. These include artistic – artisti, artist, art – historical – ancient, indigen – architectural – architecitur, demolit – and colonial aspects – coloni, colon. Hence, the topic is also interpreted to draw on words from a vocabulary of culture, and here we could add words referring to space, giving it the label *Space/Culture*.

The most probable words to appear in the abstracts that focus on the *Ethnic/Migration* topic appear to share a focus on migration – immigr, young, migrat, migrant – and aspects of ethnicity – ethnic, integr, cultur. This interpretation seems supported by the words generated by the FREQ, Life, and Score algorithms, which pick up almost the same set of words and, in addition, words focusing on presumably key aspects and potential social problems related to diversity – multiethn, religios, countercultur, segreg, integr, racialis – as well as specific ethnic identities – iran, turkish. This is why this topic has been given the name *Ethnic/Migration*. Again, as was the case for the previous two topics, this topic is not easily comprehended through the lens of a prevailing dichotomy as presented in the theory chapter. The analyses of full-text dissertations that will be in focus in the next two chapters may offer us some guidance in this respect.

---

104 Particularly, but far from exclusively, a good share of the dissertation abstracts scoring the highest could be fitted under the umbrella of urban studies, with its focus on space, community, and case studies within an urban context not seldom considered with crime. With this said, we have to keep in mind that dissertation abstracts that are not at all examples of urban studies but instead, for instance, historical studies, environmental studies, and criminology, seem to give high scores for this topic from the topic model.
In *Class/Inequality* we find the fifth least prevalent topic in which the word class has by far the highest β or probability of occurring within the topic. This word is followed by other occurrences that in this context are interpreted as encapsulating structural aspects of social inequalities (*structure/agency*)—age, mobil, disabl, incom, structur. Besides picking up many of the same words, the FREX and Lift also pick words that could be sorted under the umbrella of social inequalities and class, like those associated with income levels—household, pension—or what is interpreted as critical distinctions of cultural practices found in, for instance, different sports—sport, inferior, golf, highbrow, marxian, soccer. Thus, given the two most prevalent aspects of the topic, class and inequalities, it is called *Class/Inequality*.

Similar to the previous topic description, *Health/Demography* is, following its β levels, particularly dominated by one word, health, which is followed by words that seem to fall under the umbrella of themes related to health from a sociological point of view—mortal, suicid, status, risk, childhood, class. Second, there is a set of tokens seen to adhere to words associated with quantitative methodology (*quantitative/qualitative*)—factor, condit, associ, rate, population. The FREX, the Lift, and the Score partly present a repetition of the words already mentioned but here we also find other health themes besides suicide—cardiovascular, smoke—and an indication of more specific quantitative studies of censuses targeting large human populations on the national level (*macro/micro*) or, put simply, demography—socio-demograph, russia, census-link.\(^{105}\) Thus, this topic of *Health/Demography* is seen to be associated with the *quantitative community*.

According to the distribution of β values, the third least prevalent topic in this abstract corpus, *Domestic/Narrative*, appears to be relatively 'flat' in that not a single or two specific tokens are dominant in terms of having significantly word probabilities than the rest of the top words. The meaning of the most probable words also seems to be a bit wider than the other topics. The interpretation suggested is that we have words that, on the

\(^{105}\) Note that while the topic is labeled *Health/Demography*, dissertation abstracts that use the more generic words for quantitative methods would probably show decent scores for this topic. However, when reading through the abstracts with the highest *Health/Demography* score, it became clear that almost all dissertations were studies of the socio-demography type that investigate health differences with the help of censuses. This is, of course, one of the reasons why the discussion chapter will argue for the need to combine machine learning techniques with theoretical informed interpretations. This dissertation aims, after all, to explore what computational text analyses of dissertation abstracts and full-text dissertations in sociology suggest about the sociological imagination and the paradigms in sociology to which the theory chapter alludes.
one hand, refer to presumably problematic aspects within the domestic realm — house, office, victim — and, on the other, words echoing qualitative methods (quantitative/qualitative) and the humanities (science/humanities) — interview, mean, situate, narrate. Looking at the combination of the latter set of words in the light of the FREX, the Lift, and the Score, picking up words such as story, fieldnot, and, once again, narrate, it appears to surround the theme of interpreting the meaning of narratives based on qualitative methods. Thus, this topic of the abstract corpus is referred to as Domestic/Narrative to hold both overarching themes under one label, and is interpreted to be associated with the qualitative community.

The label given to the second least prevalent topic in the corpus is Education/Field, which is grounded on the fact that the word educ, as in the word stem education, scores the highest for all topics and three times more than the second most probable word in this topic. Further, the words that follow have connotations to education — student, choice, school, high — in combination with what is interpreted as the common field theory approach based on forms of capital — capital, field. The other algorithms primarily pick up the same words for education with a few additions associated with educational stages — junior, upper, secondary, post-secondary. While the most prevalent tokens appearing in the topic of Education/Field are thematically specific, they do not seem to have a clear relation to any of the nine dichotomies raised in the theory chapter.

When comparing the topical distribution of the corpus, the least likely topic to show up in a dissertation abstract is referred to as Problems/Treatment. Looking at the most probable word occurrences in this topic, they appear to fall under the category of social problems and their prevention — violence, alcohol, drink, treatment, client. Further, we have the two tokens effect and control, which can be interpreted in the realm of treating these social problems or, perhaps, refer to words like (measuring) effects and control (variable) used in quantitative methodology (quantitative/qualitative). Looking at the output of the other algorithms suggests the same words with some additions within what seems to be the same ballpark of social problems — drinker, homicid, beverag, epidem, delinqu, outpati. Given the good share of evidence for references to social problems and their treatment, as well as the ambiguity in relation to the quantitative/qualitative dichotomy, the topic is named Problems/Treatment.

With an individual interpretation of the 20 topics in place, we now have a general idea of the key thematic streams running throughout the abstract corpus, as well as their relative prevalence. While a decent part of the topics appeared to be strongly associated with one or a few of the sociological dichotomies presented in the theory chapter, there was a large share that
seemed to capture thematical arrangements that were instead disassociated from any dichotomy. Thus, at this stage of the analysis, based on the computational text analysis tool known as topic modeling, the evidence for an overarching division in the corpus between a *qualitative community* and a *quantitative community* seems somewhat scarce. However, this section has only scrutinized the topics one by one; so to investigate if there are wider relationships running between the topics or even communities of topics in the abstract corpus, the next section will investigate how often topics appear together in the corpus by computing topic correlations and clusters.

**Topic Correlations and Communities**

In this section, we will look into correlations between topics, that is the aggregated value of how often the topics tend to co-occur with one other within the abstracts for dissertations in sociology in focus here. On top of topic correlations, it is further possible to investigate if there are one, two, or more distinct communities of topics based on clustering edge betweenness. Figure 9 is a visualization of such an inquiry generated with these computational methods. In it, edges indicate correlations, the strength of which is visualized by the thickness of the lines, the 20 nodes represent the 20 topics, the sizes of which indicate the topics’ prevalence in the abstract corpus, and the colored fields indicate that the topics positioned therein are part of the same community. In the figure, there are four communities: one colored in light red constituted by nine orange nodes, another in light green based on nine blue nodes, and, lastly, two single-topic communities in light purple with a yellow node, and light blue with a green node, respectively.

\[^{106}\text{The topic correlations are the maximum a posteriori probability (MAP) estimates for the topic proportions theta value (θ), providing the marginal correlation of the mode of the variational distribution (see https://rdrr.io/cran/stm/man/topicCorr.html [accessed 2022-11-29]).}\]

\[^{107}\text{Edge betweenness algorithm treats the edges as distances and measures the shortest path between each pair of nodes, in this case the topics (see https://igraph.org/r/doc/cluster_edge_betweenness.html [accessed 2022-11-29]).}\]

\[^{108}\text{In other words, the larger the node, the higher it is positioned in Figure 9.}\]
Figure 9. Network representation of correlations between the 20 topics computed based on how likely topics are to co-occur within the same abstracts. A modularity was added to demark communities, hence the different colors. The size of the nodes indicates how frequently the topics occur in the corpus, whereas the thickness of the edges connecting them indicates how strongly the topics are correlated.

Looking at Figure 9, probably the first thing one notices are the two larger communities visualized as two separately colored fields in light red and light green. Before diving deeper into these communities, however, we can also see two smaller communities consisting of only a single topic that is not part of the larger network, in light blue and light purple, respectively. These two topics, *Space/Culture* – interpreted to encapsulate words related to spatial and cultural matters – and *Domestic/Narrative* – construed to pick up words related to meaning-making, stories of different kinds, and words associated with the domestic realm like housing – are not correlated with any other topic. This means that if a dissertation abstract uses words associated with one of these topics, we are given no clue about what other topic they also address out of the other 19 topics explored here since there is no clear correlation.
In contrast to the two tiny communities, we can draw out plenty of indications of how different topics relate to each other from the two larger colored fields. Again, it is worth repeating that shared color indicates a community in the sense that these topics tend to correlate in the dissertation abstracts. Knowing that the strength of each topic-to-topic relation is visualized through the thickness of the edge running between both topics, we can see that not all topics within the two larger communities are correlated, and the ones that do share an edge are not equally strongly correlated. For instance, in the light red community, the topic Micro/Humanist is correlated with three topics, the strongest edges going to, in descending order, the topics Theory/Knowledge, Discourse/Politics, and Qualitative. Hence, based on this method, we can assume, among other things, that in sociology dissertation abstracts that draw on a vocabulary of words related to the micro level of society or the humanities, it is quite probable that we would also find words related to sociological theory or different types of knowledge. Throughout the light red community, most topics correlate with either two or three topics, yet Service/Profession is only linked to one other topic and Qualitative to five topics.

Turning to the light green community, there are topics with one to five edges. On the low end, we find, for example, the topic Welfare/Countries that is only linked with Positivist/Risk, whereas the topic Health/Demography is on the high end with as many as five edges, yet all these edges are weaker than the one running between Welfare/Countries and Positivist/Risk. Therefore, it appears that if one of the dissertation abstracts uses words picked up by the topic Welfare/Countries – i.e., words seemingly related to the theme of welfare state policy and cross-country comparisons – there is a reasonable chance that we will also find words constituting the topic Positivist/Risk in that document. In comparison, if a dissertation abstract draws on words located within the Health/Demography topic, it is less probable that we will also find words associated with another specific topic. Rather, abstracts associated with the topic Health/Demography are associated with five different topics but to a considerably lesser extent: Positivist/Risk, School/Media, Class/Inequality, Problems/Treatment, and Education/Field.

When appraising the topics constituting the two larger communities and comparing them to each other, a thematic arrangement seems to emerge that resonates with the findings from the previous analytical chapter on word constellations. Thus, in relation to the dichotomies presented in the theory chapter and the manner in which these are guiding interpretations of the visualizations presented here, we can see that most interconnected nodes of the community in light red are represented by the topics Qualitative (quantitative/qualitative, macro/micro, and empirical/theoretical), Micro/Humanist (macro/micro, science/humanities, and holism/individualism), and Histo-
ry/Economy (science/humanities). Added together, these three topics are interpreted as picking words associated with multiple poles resonating with the qualitative community. In this community, we further find Theory/Knowledge, constituted by words with connotation to the theoretical realm (empirical/theoretical) and studies of knowledge (as well as, to a lesser extent, holism/individualism, structure/agency, and macro/micro), and Discourse/Politics, as in words related to discourse analysis and political questions (interpreted as leaning towards explanatory/interpretative and macro/micro). These five topics seem reasonable to appear together as they are, each in their way, typical for this side of the so-called methodological divide. The exceptions are the words resonating with political questions and economy, suggesting that these themes can be addressed in abstracts that also address words associated with the qualitative community.

Further, there are three topics, all of which are not clearly associated with any prevailing dichotomy but which are all linked to the topic Qualitative, which gives us a clue that these seemingly empirical themes can be associated with words for qualitative methods. These topics are Service/Profession, with its words related to social work and medical professions, Organization/Movement, interpreted as holding words echoing organization studies and social movements, and Ethnic/Migration, constituting a vocabulary of words relating to ethnicities, migrants, and (cultural) integration. Lastly, the fifth topic linked to Qualitative is Labor/Market – drawing on words associated with concepts like work, labor, and types of markets – which has a unique position in Figure 9 since it forms a bridge between the two larger communities with its link to the topic Positivist/Risk. This suggests that the words constituting Labor/Market are not exclusively related to the macro level (macro/micro) as was partly suggested in the previous chapter, but might rather be associated with both sides of the methodological divide.

As has already been suggested, the light green community to the south is interpreted as representing the quantitative community that is opposed to the qualitative community. The most seemingly central topics, which are interconnected, are here Positivist/Risk (which can be interpreted as examples of the following dichotomies: science/humanities, explanatory/interpretative, quantitative/qualitative, empirical/theoretical) and Health/Demography (which are in turn examples of the following dichotomies: macro/micro, quantitative/qualitative, and empirical/theoretical). Both of these topics go well with quantitative sociology with its inclination towards words associated with scientistic or positivist thinking and comparative studies of larger populations of people. The latter topic, Health/Demography, is linked to another topic constituted by words indicating different aspects related to quantitative studies of people, namely Family/Background, which pick up words for social backgrounds like occupations and family roles (interpreted
as associated with **structure/agency, macro/micro, quantititative/qualitative, and empirical/theoretical**. In addition, Figure 9 shows that the topic *Health/Demography* also correlates with *Class/Inequality*, another topic associated with social inequalities – this time in terms of social structures based on class, age, income, and more (**structure/agency**) – as well as two topics interpreted as not being coupled with any of the sociological dichotomies: *Problems/Treatment*, associated with social problems like alcohol consumption, violence, and measuring the effects of potential remedies, and *Education/Field*, drawing on words related to education and theories of social fields.

The other central topic in this community, *Positivist/Risk*, is also linked to *Education/Field* as well as the three remaining topics, which can give us hints about some of the more specific empirical topics associated with the *quantitative community*. To the left, there is a link leading to *Welfare/Countries*, which picks up words associated with the welfare state, policy, and what appear to be comparative studies of different countries on the macro level (**macro/micro**). Based on the knowledge we have gained from the previous empirical chapter and the visualization hereby presented in Figure 9, it is expected that this topic would appear in the quantitative sociology community. On the other side, to the right, another link goes to a topic that was interpreted as not being associated with a prevailing dichotomy, namely *School/Media*, constituted by words associated with schools, media outlets, and debatable subjects. Lastly, *Positivist/Risk* is connected with *Labor/Market*, which serves as a bridge between the two communities. Recalling the results from the previous empirical chapter, this is not unexpected since the labor market was a word constellation constantly falling on the quantitative side of the methodological divide.

From this analysis of how the 20 topics correlate with one another in this abstract corpus, and the communities they form through measuring edge betweenness, we have mapped out the degree to which different topics relate to each other, which topics are most central for each community, and, perhaps most importantly, have seen an overarching thematic arrangement emerge from the two major communities that can be used to roughly represent the abstracts. The latter can be summarized as comprising what appears to be a thematic representation of the methodological divide revealed in the previous chapter, with the *qualitative community* on one side and the *quantitative community* on the other. Each side appears to be constituted by some topics that are related to either set of poles in the dichotomies taken to be characteristic of this methodological divide, which in the previous analytical chapter was taken to be primarily the methodological dichotomies (**science/humanities, explanatory/interpretative, and quantitative/qualitative**), but here we might also sense some of the ontological ones (**ho-**
lism/individualism, structure/agency, and macro/micro), and, given the position of the Theory/Knowledge chapter, perhaps also the epistemological dichotomy empirical/theoretical. Worth noting is that there were a good set of topics on both sides that seem to express specific empirical themes that are not automatically associated with this divide. This is why it is hereby suggested that the latter entails that these empirical words have a stronger probability to be used together with words specific for one side of the methodological divide or, to use a more graphic language, these empirical themes are entangled with either a vocabulary resonating with the quantitative community or a terminology echoing the qualitative community.

With the results found in this analysis, we can now continue to answer the research questions by looking at the contingency of the thematic arrangements exposed in this section, interpreted as primarily separating topics associated with the qualitative community and the quantitative community. This is done by investigating the effect of time and place across the topics. As in the previous chapter, we will start with the former and end with the latter.

The Effect of Time on the Topic Proportions

In this section, we will look at the impact of time on the distribution of the topics over the corpus. This is done by leveraging the structural topic model (STM) framework – where the effect of factors can be measured on the Latent Dirichlet Allocation (LDA) – so that expected topic proportions can be computed for the time period studied, 1980-2019. As will be the case for all analytical chapters in this dissertation, the variable of time is here taken from the defense year of the dissertations. This is done by computing a regression where theta (θ) values or topic proportions are the outcome variable, giving us the conditional expectation of topic prevalence based on the content of the abstracts.¹⁰⁹

¹⁰⁹ In other words, in this regression, the abstracts are the units, the outcome is the topic proportions for each abstract about a topic in the model, and the covariates are the time variable. This method incorporates uncertainty by the method of composition (see https://www.rdocumentation.org/packages/stm versions/1.3.6/topics/estimateEffect [accessed 2022-11-29]).
Figure 10. Temporal proportions of the 20 topics sorted by their relative prevalence in the abstract corpus. The abstracts have been sorted into corpora by the demi-decade in which the dissertations were defended.

For this analysis, we want to keep in mind the analysis of time presented in the previous chapter, based on word correlation networks and cluster analysis of abstracts used in sociology dissertations. It was there suggested that while it was easy to find references to specific empirical words associated with different time eras in the abstract corpus, it seemed harder to map out temporal patterns based on words associated with specific methodologies or epistemologies. Put more persuasively, the chapter on word constellations led to the interpretation that a methodological divide is not unevenly distributed over time in the corpus but rather appears to recur throughout the time period. Thus, it is now of interest to see if this pattern is sustained or challenged when topics modeled on the dissertation abstracts and their correlations form the basis of the analysis. The section is based on a structure where a general visualization of all topics is first displayed and discussed. Based on this analysis, three categories of topics are targeted and scrutinized category by category.
In Figure 10, we can see how the 20 topics are distributed over the period of 40 years (1980-2019) separated by demi-decade (five years), which make up a time sequence constituted by eight elements. The figure is based on relative measurement so that each element or demi-decade adds up to 1.00 or 100% topic proportion distributed over the 20 topics. The topics are sorted after their prevalence in the corpus – the same manner and color scheme as in Figure 8 found in the previous section. Thus, the most prevalent topic, Theory/Knowledge, is positioned at the top of the figure in white, while the least prevalent topic, Problems/Treatment, is to be found at the bottom in brown. With these temporal proportions of the 20 topics at hand, we are given more nuances to each topic’s prevalence in this abstract corpus. For instance, the mentioned topic Theory/Knowledge seems relatively popular throughout the timeframe of the data but, according to the figure, is one of the least probable topics to occur in the analyzed abstracts coupled with dissertations defended in the period 2015-2019, and not particularly popular in 2010-2014. Thus, the figure can provide us with clues about how the topics evolve over time in terms of their popularity across the eight demi-decades. For instance, the topics of History/Economy and School/Media appear to show a steady decrease in relative popularity over time in these abstracts. In contrast, the topics of Gender and Discourse/Politics seem to be increasingly prevalent over time. By investigating the 20 topics’ trajectories and comparing them to each other, four exclusive categories have been worked out. These categories are topics that seem to have a rather balanced prevalence when considering the whole time period or that are either most prevalent during the 20th century or the 21st century.

In Figure 11, we find seven time series that express the expected topic proportion for each year in the time period 1980-2019. In other words, the time series visualizes how probable it is that a topic would appear in the dissertation abstracts categorized by year of defense. The horizontal x-axis represents time sorted by full years and the vertical y-axis the topics’ expected proportion over the corpus at a given time. The line shows the mean of the topic for each year and the colored field represents the prediction interval (i.e., the mean’s variability based on the spread of the individual observation at that year). Hence, for instance, a broad colored field spanning a period of years would entail that the dissertation abstracts hereby analyzed showed a wide range of low and high topic probabilities for those years.

110 It is important to recall that the number of sociology dissertation abstracts comprised in the periods is not even. For comparison, the first three periods – 1980-1984, 1985-1989, and 1990-1994 – make up the smallest periods with only 58-63 abstracts each, while the fifth period, 2000-2004, is the largest with 160 abstracts.
Figure 11. Time series showing the prevalence of topics interpreted to have a rather balanced prevalence in abstracts for dissertations defended on either side of the second millennium. The color range shows the confidence interval of each topic (95%).

The time series in Figure 11 show a relatively stable annual distribution of expected topic proportions over time. Put in simple terms, the topics visualized in the figure seem not to be typical for a certain era but rather recur relatively balanced throughout the time period 1980-2019. While each of the time series has its own particularities, we can sort their recurrence over time by a continuum from stability to fluctuation. On the former end, we find two of the topics in this abstract corpus, Positivist/Risk and Qualitative, that are interpreted to be particularly stable in the sense that time appears to have little effect on how the topics are distributed in the abstract corpus. The lines of these two topics show a horizontal pattern that remains within a narrow span of expected topic proportion values throughout the observed time period – where Qualitative is occurring slightly more frequently than Positivist/Risk throughout the whole time period. These topics can be interpreted to represent different poles in several of the dichotomies interpreted to represent the methodological divide, for Qualitative the qualitative and the micro poles, and for Positivist/Risk the science, explanatory, and quantitative poles. If we put the more empirical words coupled with the two topics in parenthesis, which have a low probability to occur in the topics, it would seem to make sense that these topics would be stable over time. This is because their most probable words refer to general words used in empirical studies within sociology, which presumably occurs to some degree in most sociology dissertation abstracts.

On the other end, we can place topics like Service/Profession, Welfare/Countries, and, in particular, Labor/Market, which share the pattern of a
fluctuating expected topic proportion that peaks and drops at an interval of about a decade. In contrast to the stable topics, these fluctuating topics all comprise words that can be associated with certain aspects of the social world. Since one could assume that far from all abstracts for sociology dissertations would refer to these specific words, it is understandable that these topics do not recur steadily over time. The same goes for the two remaining topics that characterize this abstract corpus, *Ethnic/Migration* and *Problems/Treatment*, which are a little less fluctuating but with a marked peak around the turn of the century. We can only speculate what the reason for a strong fluctuating pattern might be – such as if it reflects an intermittent intake of doctoral candidates at different universities or not – but this matter goes outside the scope of the research questions informing this chapter.

Recalling the analysis put forward in the previous chapter, it seems interesting to explore the popularity of the 20 topics over time in relation to their given position in either of the two communities. By inferring the results drawn from Figure 10 – where a network analysis of correlations between topics generated two communities that were interpreted as illustrating a methodological divide – we find that the topics presented in Figure 11 form an equilibrium in relation to the *quantitative community* and the *qualitative community* that the dichotomies informing the interpretation of results in this dissertation allude to (as presented in the theory chapter and alluded to as one of the results that the previous chapter pointed toward). This is because three topics are positioned in the former community – *Positivist/Risk, Problems/Treatment*, and *Welfare/Countries* – and three in the latter – *Service/Profession, Ethnic/Migration*, and *Qualitative*. The last topic, *Labor/Market*, was interpreted as a bridge between the communities since it is linked to both, according to the network analysis performed. Thus, in light of the topics’ position in Figure 10, the *quantitative community* and the *qualitative community* seem to be balanced in Figure 11 since both sides are represented by an equal number of topics.
Figure 12. Time series showing the prevalence of topics interpreted to lean towards abstracts for dissertations defended during the 20th century. The color range shows the confidence interval of each topic (95%).

In Figure 12, there are six different time series, each representing a topic that is characterized by having its most popular period, in terms of the abstracts for sociology dissertations defended, in the 20th century rather than the 21st century. As is clear, the topics are not equally popular in the earlier century. Two of the topics demonstrate a particularly strong leaning towards the 20th century, namely Class/Inequality and, even more so, History/Economy. The topic of History/Economy shows the strongest descending trend, from being the most prevalent topic throughout the 1980s to sequentially dropping down two-thirds at the beginning of the 21st century and forward. In comparison, the topic hereby labelled Class/Inequality displays a similar trend with its peak in the late 1980s but is less popular in the corpus, and its drop over time appears a little less dramatic with a small trend break in the late 2010s. Thus, it seems possible to propose that words relating to social inequalities of different sorts, in particular social class, as well as words echoing historical and economic concepts, were more popular in abstracts for sociology dissertations defended in the 20th century than in the 21st century – and predominantly so in the 1980s.

Two other topics, School/Media and Space/Culture, also have a substantial peak in the abstracts for dissertations defended in the 1980s but here the downward trend for the rest of the time period appears less clear. Both topics express small bumps in the 21st century, which do not, however, exceed the 1980s peak. In other words, it seems as though words associated with matters relating to school, media, space, and culture are prominent in the abstracts stemming from the 1980s but have one or two increase and decrease instances in the 2000-2010s. The two last topics, Organization/Movement and Theory/Knowledge, seem even more equally distributed in the abstract
corpus in terms of time. While both topics fluctuate throughout the whole time period, it is, however, clear that they show a stronger overall prevalence in the 20th century than in the 21st century. Thus, while words connected to organizations, social movements, theory, and knowledge appear from 1980 to 2019, they show stronger prevalence in the 20th century. This trend is clearer for the latter topic, Theory/Knowledge, since it has its highest point in the 1980s and thereafter an overall downward trend, yet this trend is subtle, with a recovery in the mid-1990s as well as throughout the 2000s, and then dropping in the 2010s. Nevertheless, it is also worth noting that the prediction interval (represented by the width of the colored area) is the widest for Theory/Knowledge, which suggests that this topic has a dispersed prevalence in this abstract corpus, with a large width between observations with a high and low expected topic proportion. Drawing on the insights gained from the previous chapter, this makes sense since words appealing to theory seem to be rather general for all dissertation abstracts, while words related to knowledge appear to be more specific for abstracts for dissertations that seem to be about the sociology of knowledge and equivalent fields that tend to draw on a theory of science vocabulary. This too will be explored further once the analyses of the full-text dissertation corpora is presented in the two chapters that follow. For now, it is just noted that this could be a plausible interpretation.

Let us again try to apply the topics’ classification vis-à-vis the methodological divide revealed in the previous section, but now to the trends suggested in Figure 12. The two topics in this abstract corpus with the strongest leaning towards the 20th century, Class/Inequality and History/Economy, were interlinked with the quantitative community and the qualitative community respectively. Further, there is one topic that was positioned on the quantitative side of the topic correlation network, School/Media, and another, Space/Culture, that is neutral in the sense that it did not correlate with either community, both of which are a little less attuned to the 20th century but that had their peak in the 1980s. Lastly, the topics with a good probability to appear throughout the time period but with a stronger prevalence in the 20th century, Organization/Movement and Theory/Knowledge, were both found to adhere to the qualitative community. Thus, in relation to the methodological divide, which is based on a set of dichotomies presented in the theory chapter – science/humanities, explanatory/interpretative, quantitative/qualitative, and macro/micro – and guiding the interpretation of results generated in this chapter since it was found in the previous chapter, we seem to have a rather good balance with perhaps a small tilt towards the qualitative side in the topics interpreted to be associated with the 20th century rather than the 21st century.
Figure 13. Time series showing the prevalence of topics interpreted to lean towards abstracts for dissertations defended during the 21st century. The color range shows the confidence interval of each topic (95%).

Figure 13 is constituted by seven topics that all have their peak in the 21st century instead of the 20th century, thus displaying an overall upward trend although of various extents. The most dramatic 21st-century pattern can be found in the fluctuating topic Gender. The topic shows a flat period in the 1990s and two higher peaks in the years right after 2000 and 2010, respectively – both representing the two highest expected topic proportions for the 21st century period for all topics – followed by a drop for the period 2015-2019. One explanation could be the two-sided quality of Gender, both picking up generic words for gender identities and roles as well as more specific terms related to feminist thinking; we would expect the former to show up rather steadily throughout the time period and the latter to make its grand appearance at the turn of the 21st century, given the rise of feminist theory and gender studies. Perhaps the drop at the end of the time series suggests that this specific field is losing its foothold in sociology dissertations, the reasons for which – such as the introduction of gender studies as a research subject in its own right, with its own dissertations, in Sweden – are, however, outside the scope of this study.

Thereafter, we have three topics – Discourse/Politics, Domestic/Narrative, and Micro/Humanist – that all show a stable upward trend. The strongest trend can be seen in Discourse/Politics, which seems to have its low point in the 1980s and a high point in the 2010s, a sturdy increase disrupted only by a small dip around 2005. Domestic/Narrative and Micro/Humanist seem to have a similar trend to Discourse/Politics but, in comparison, they are displaying a more horizontal upward trend at a lower amplitude and with less clear ‘bumps’. Thus, words related to discourses, politics, domestic matters,
and narratives, as well as words associated with the micro level of society and a humanist research tradition, seem to have been increasing in frequency steadily across time in the corpus. As will be explained further below, this tendency could perhaps be explained by the fact that the three topics share the characteristic of picking up somewhat generic terms associated with the *qualitative community*, where the more isolated position of the topic *Domestic/Narrative* calls for further clarification.

Lastly, the topics *Education/Field, Family/Background,* and *Health/Demography* do not express an upward trend from the 20th to the 21st century. Rather, these time series express a variable prevalence over the time period studied, yet have their all-time peak positioned somewhere around the mid-2010s. While *Education/Field* manages to sustain its increased expected topic proportion throughout the 2010s, *Health/Demography* returns to its normal levels in the late 2010s. Regarding *Family/Background,* there are two peaks, with the additional one taking place in the early 2000s. The three topics make up a mix of both general words for quantitative methods and social categories of people, as is seen partly in both *Health/Demography* and *Family/Background,* and specific empirical topics within sociology like words related to the domains of health, education, and families.

When looking at these seven topics strictly through the communities suggested in the analysis of topic correlations above, we find that the topic with the highest expected proportion points in the 21st century, *Gender,* is positioned in the *quantitative community.* In contrast, the three topics with a straight upward trend – *Discourse/Politics, Domestic/Narrative,* and *Micro/Humanist* – were all part of the *qualitative community.* The remaining topics, which all showed a rather balanced prevalence throughout the time period but clearly peaked around the 2010s, are represented by three topics associated with the *quantitative community: Education/Field, Family/Background,* and *Health/Demography.* Therefore, the two communities appear to almost be in equilibrium through 1980-2019 but with a bent towards the quantitative side, which is sustained by the patterns exposed in the previous figure showing topics with a leaning toward the 20th century.

To conclude this analytical section on the effect of time on the thematical arrangements in the abstract corpus, the presented interpretation of the structural topic model has suggested a set of topics that seem to recur rather consistently over time, another set that appear to have had their heydays in the 20th century, and a third interpreted to reach their peak in the 21st century. In relation to the methodological divide, it appears as if topics’ positions in the *qualitative community* and the *quantitative community,* respectively, are fairly equally distributed in the three categories of time (i.e., the 20th century, the 21st century, and balancing between the centuries). However, there is
perhaps some evidence in the direction of a small decline in the period 1980-
2019 for the topics associated with the \textit{qualitative community} in favor of
those falling within the \textit{quantitative community}. Nevertheless, thus far this
analysis of thematical arrangements appears to be in line with the analysis of
word correlations presented in the previous chapter. It therefore seems as if
we can tentatively assume that the methodological divide is rather consistent
over time in this corpus of abstracts for sociology dissertations.

The Effect of Place on the Topic Proportions

We will now turn to investigate the influence of place on the topics modeled
on the abstract corpus. The procedure is the same as was described in the
previous section except that the factor of time is exchanged for the factor of
place, represented by the university where the dissertation was defended.
Again, we want to recall the results generated in the previous chapter, where
word correlation networks were analyzed and word constellations mapped
out. In stark contrast to time, which seemed to have little impact on the over-
all pattern alluding to a divide between a \textit{qualitative community} and a \textit{quan-
titative community} in the corpus, place appeared to be a strong factor. The
analysis indicated that abstracts stemming from different universities were
associated with words on one of the two sides of the methodological divide
as mapped out in the data, as well as some words associated with particular
empirical themes. Since the analyses of time put forth in this chapter and the
previous chapter implied the same overarching pattern – i.e., that the meth-
odological divide recurs over time but that some empirical themes come and
go while others recur as common references – it is now relevant to discover
if the same holds when analyzing the influence of place (i.e., where the ab-
stract being analyzed comes from) on this corpus. In other words, is place
still a strong factor for differentiating the two communities alluded to earlier
(i.e., \textit{quantitative} and \textit{qualitative}) when scrutinizing the topics instead of
word correlations? The section follows closely the structure of the previous
section, where a general visualization of all topics is initially examined and
then broken into smaller categorizations of topics that are analyzed separate-
ly.
Figure 14 is a visualization of the effect that the covariate of place, taken from the metadata of the universities where the dissertations were defended, has on the expected proportion of the 20 topics over the abstract corpus. The topics are again sorted by their prevalence and assigned the same colors as in the previous section. To begin with, we can see that the expected topic proportions recorded are neither equally spread between the 20 topics nor the five universities from which the abstracts stem. In other words, the topics of the abstract corpus seem to be unequally coupled with the universities, which here serve as different representations of place. Looking at the pattern emerging within the figure, it seems that the topics are in fact associated with abstracts from a certain university or universities. For instance, we can see that the most prevalent topics for abstracts originating from Stockholm University are Health/Demography and thereafter Family/Background, while the equivalent topics for Umeå University are Welfare/Countries and Positivist/Risk. These four topics all appeared in the quantitative community of abstracts found earlier in this chapter. In contrast, abstracts for sociology dissertations defended at Uppsala University evidently display the highest probability to draw on the topic Theory/Knowledge followed by Hist...
ry/Economy, and, to a slightly lesser degree, the most probable topics for abstracts stemming from Gothenburg University are Service/Profession and Qualitative. A similar selection is, however, harder to perform for the abstracts of sociology dissertations defended at Lund University. The topics of abstracts stemming from Lund appear to be, in descending order (with regards to prevalence), History/Economy, Theory/Knowledge, and School/Media, but it is only the latter where abstracts from Lund have the highest topic proportion compared to the other universities. Nevertheless, the top topics for abstracts stemming from Uppsala, Gothenburg, and Lund seem to belong to the qualitative community when we use the theoretical lens that the dichotomies presented in the theory chapter offer us.

In contrast to the previous chapter, where abstracts originating from Gothenburg were interpreted as being positioned in the middle of the so-called methodological divide, the analysis presented in this chapter suggests instead that these abstracts belong to the qualitative community, together with those stemming from sociology dissertations from Uppsala and Lund. Thus, abstracts from these three universities seem to form a group of abstracts with a qualitative leaning that can be compared to another with a quantitative leaning, constituted by those for dissertations defended in Stockholm and Umeå. So, based on the tendencies discussed in relation to Figure 14, it seems as though we have three categories of topics to investigate further to see if the pattern holds for all topics. The first two categories represent topics for abstracts stemming from Stockholm or Umeå, on the one hand, and topics for abstracts for dissertations defended in Uppsala, Lund, and Gothenburg, on the other. The last category is, instead, topics that are fairly equally distributed over the abstract corpora of the universities in both communities, that is topics not associated with either a single university or a particular side of this divide. Thus, let us now look at the topics of this abstract corpus through the lens of the universities they stem from.
Figure 15. Topics where the factor of Stockholm or Umeå is interpreted to have the strongest effect on the topics’ prevalence in the abstract corpus. A topic is represented by a bar chart in which the university covariates make up the bars. The expected topic proportions are given by the gamma of the model.

Figure 15 displays the expected topic proportion for six topics over the universities. Put differently, the figure shows the effect that factors of spatial representations have on a group of topics. Each topic is represented by a bar chart in which the five universities are denoted by a bar that represents the mean value of the university covariate. To recall, the selection of topics is here based on singling out topics where abstracts from Stockholm or Umeå expressed the highest values on the y-axis. The purpose of the figure is to investigate the strength of the association between the topics and the universities, as well as if there is a pattern to what kind of topics are associated with these two universities.

One of the first things to stress when interpreting Figure 15 is that all six topics suggested to belong to the quantitative community are present. This includes the two topics found to be most central for the community and the
ones most clearly associated with words referencing the ‘right’ poles of its key dichotomies (i.e., science/humanities, explanatory/interpretative, quantitative/qualitative, and macro/micro), namely Positivist/Risk and Health/Demography. Another thing to notice is that the expected topic proportions between the universities vary considerably from topic to topic. The interpretation put forth is that the most convincing results are found in the topics Health/Demography and Welfare/Countries, since these show high expected topic proportions for abstracts for dissertations defended in both Stockholm and Umeå, where the abstracts from one of the universities express the highest value and the other university the second highest. This pattern is repeated for the topic Positivist/Risk, with the exception that the expected topic proportion frequencies for all topics are lower. Thereafter, we find two topics, Family/Background and, with slightly lesser expected topic proportions, Education/Field, that are characterized by abstracts from Stockholm expressing the highest effect followed by abstracts from Gothenburg and Umeå in third place – even a shared third place for Education/Field. Lastly, the least substantial result is found in the topic Problems/Treatment. Here, abstracts from Stockholm have the highest expected topic proportion and those from Umeå the lowest.
Figure 16. Topics where the factor of Lund, Uppsala, or Gothenburg is interpreted to have the strongest effect on the topics’ prevalence of in the abstract corpus. A topic is represented by a bar chart in which the university covariates make up the bars. The expected topic proportions are given by the gamma of the model.

Figure 16 shows the expected proportions for eight other topics where, in contrast to the previous figure, abstracts from Lund, Uppsala, or Gothenburg show the strongest effect of place on the topics. Again, the figure shows the expected topic proportions (y-axis) over the five universities (x-axis). The first thing to note with the eight topics is that six of them take part in the qualitative community mapped out in Figure 9, whereas the other two were
isolated from both communities in that figure. Figure 16 further includes all of the topics that were interpreted to be most clearly associated with qualitative sociology since they pick up words related to the opposite poles of the dichotomies presented for the six topics above (i.e., science/humanities, explanatory/interpretative, quantitative/qualitative, and macro/micro), namely Micro/Humanist, Qualitative, Discourse/Politics, and History/Economy. Yet, as we will again see, the expected topic proportions of the eight topics are far from evenly distributed, and not only associated with the three universities that are suggested to have produced abstracts that lean towards the qualitative community.

The strongest case for an overarching methodological division between the two groups of abstracts stemming from universities – Stockholm and Umeå, on the one side, and Uppsala, Lund, and Gothenburg, on the other – is seen in the topic Theory/Knowledge. In this topic, abstracts from Uppsala display the greatest and Stockholm the lowest expected topic proportion at a high level of contrast. Further, abstracts from Lund and Gothenburg score above Umeå. The same pattern but at a lower amplitude is found in the topic History/Economy, as well as in Domestic/Narrative, the latter with abstracts from Lund instead of Uppsala recording the top relative effect. An interesting aspect of this result is that Domestic/Narrative is so clearly associated with, in descending order, abstracts stemming from Uppsala, Lund, and Gothenburg, but did not correlate with any of the correlating topics forming the qualitative community. Either we can interpret this on the lines that Domestic/Narrative shares a latent association with this community or that the topic is first and foremost associated with abstracts stemming from specific universities (Uppsala and to a lesser extent Lund) rather than a representative of the qualitative community. Looking at some of the top words constituting this topic — interview, mean, situate, narrat, stori, fieldnot — several seem to be resonating well with the key dichotomies for this community, so we will go for the former alternative with the mentioned reservation in mind.

Thereafter, we find three topics, Qualitative, Micro/Humanist, and Discourse/Politics, which are characterized by the fact that one of the three universities (Uppsala, Lund, or Gothenburg) has the highest as well as the second highest expected topic proportion in this abstract corpus, followed by the remaining of the three universities. This third university has a similar effect as Umeå or both Stockholm and Umeå. While the pattern here is a bit weaker since not all the abstracts from these three universities score the highest, these are also interpreted as typical examples of the qualitative side of the methodological divide. To begin with, the three can be located within the qualitative community shown in Figure 9, but, in addition, the three are all more or less constituted by words alluding to qualitative sociology. Qual-
itative with its qualitative method words, Micro/Humanist with words alluding to the micro level of society and humanist vocabulary, and, lastly, Discourse/Politics, which shows high probabilities for words related to discourse analysis construed as an interpretive methodology.

The remaining topics, Space/Culture and Service/Profession, show an even less clear distinction between the two groups of abstracts in relation to universities variables. In each case, abstracts for sociology dissertations defended in either Lund or Gothenburg express the highest effect, closely followed by the other two (Uppsala and Lund or Gothenburg) at the same level as Umeå or Stockholm. Thus, as was the case in the paragraph above, there is a possibility that these three topics are signifying characteristics related to place (i.e., one university) rather than a methodological divide as such. For Space/Culture and Service/Profession, this seems to be the case. Recalling the topic description given in the first section, both topics are constituted by words that are interpreted to be 'neutral' in relation to the two communities in the sense that they almost exclusively pick up empirical themes. Further, Space/Culture was an isolated node in the topic network presented in Figure 9.
Figure 17. Topics where the effect of space on the topics’ prevalence in the abstract corpus is interpreted to be either marginal or balanced between the university corpora of abstracts. A topic is represented by a bar chart in which the university covariates make up the bars. The expected topic proportions are given by the gamma of the model.

In Figure 17, which again shows the effect of place on the expected topic proportions, we find topics that were neither clearly associated with abstracts from a specific university nor positioned as one would expect concerning the two opposing communities mapped out above. First, we have a set of topics that show such marginal variances in expected topic proportion for the five universities that to make a point out these differences would be overinterpretation. Two of these three topics appear in the qualitative community in Figure 9, Ethnic/Migration and Organization/Movement, while to last topic, Class/Inequality, instead can be found in the quantitative community. Two other topics, Labor/Market and School/Media, show slightly higher values on the y-axis but should probably still be labelled marginal in the same. However, what is interesting about these seemingly marginal topics, is that they have a slightly deviating lean in comparison to the general pattern.
The topic *Labor/Market* seems to be most strongly associated with *Stockholm, Umeå, and Gothenburg*, and according to the proposed interpretation of the topic network, *Labor/Market* represents a bridge between the *qualitative community* and the *quantitative community*. This seemingly marginal reversed relation between Figure 17 and Figure 9 is also seen in the topic *School/Media*, a topic that was suggested as part of the *quantitative community* where, in descending order, abstracts from *Lund, Gothenburg, and Uppsala* show the highest values.

The most remarkable topic in Figure 17 is *Gender*, a topic that cannot be rejected as marginal when looking at its relatively high and differentiating expected topic proportions for all universities. This calls for closer scrutiny. Out of the five universities, abstracts originating from *Uppsala, Umeå, and Gothenburg* score the highest, but the values for the other two are not far behind. To recall, *Gender* was located in the *quantitative community* in the topic correlation network, but is here tilting slightly towards universities that – based on the analyses presented in this and the former chapter – seem to be more attuned to the *qualitative community*. Again, an explanation can be derived from the two-sided quality of *Gender*, with general words for gender identities often found in both qualitative studies of small populations and quantitative studies of large populations, on the one hand, and specific terms echoing a feminist and constructionist discourse associated with qualitative sociology, on the other. It was suggested in a previous analysis that the former more generic quality of the topic appears more prevalent than the latter’s more specific quality. Further, *Gender* only has one connection in Figure 9, which is to another seemingly general topic, *Family/Background*, picking up both qualitative and quantifiable aspects of identities and roles related to various concepts allied to social backgrounds. Thus, *Gender* is interpreted as a broad topic spanning both sides of the methodological divide.

To conclude this analytical section on the effect of place on the 20 topics, the presented interpretation of the structural topic modeling has suggested that the topics of this abstract corpus ought to be categorized into three sets. Two of the sets are in stark opposition to each other. The first is capturing topics interpreted as related to the *quantitative community* that seems to be associated with a certain place, comprised primarily of *Stockholm University* and *Umeå University*. In contrast, the second set of topics was primarily labeled as leaning toward the *qualitative community* and was linked to another spatial configuration based on *Uppsala University, Lund University, and Gothenburg University*. The last set of topics is construed as carrying no meaningful information vis-à-vis the methodological divide since the spatial effect was either marginal or had an explainable deviation from the general pattern exposed in the other two topic sets. Thus, concerning the methodo-
logical divide understood to be affected by place in the previous chapter on word constellations, it appears that the thematical arrangements of the abstract corpus can also be interpreted on the lines of a qualitative community and a quantitative community with an unequal distribution over place. Since all the topics of the latter community to a greater extent were associated with the same universities, we might wonder, as we did in the last chapter, if the vocabulary used by the quantitative community is more uniform than that adhering to the qualitative community.

Closing Thoughts on Thematic Arrangements

The structural topic model employed in this chapter found 20 topics constituted by relatively distinct combinations of words. While most topics were quite well distributed in the corpus, some stood out with regard to popularity. Two examples of the more prevalent topics are those that seem to pick up words related to theory, and another that appeared to be based on terms associated with gender. On the other end of the spectrum, we found two topics with almost half the score of the previous two, interpreted to be about social problems and education, respectively. By investigating how the topics correlate in the sense of appearing in the same abstracts, 18 of the topics formed a network of two major communities bridged by a topic seemingly referring to work and the labor market, whereas the remaining two topics did not reach the correlation limit with any other topic and therefore were visualized as isolated nodes.

With this brief summary of the structural topic model, we can turn to the three research questions guiding this chapter. Beginning with the first question, it investigates whether sociological knowledge appears to be fragmented. When measuring how they were affected by time, the topics can be sorted into three groupings of similar size. First, some topics seemed to be mildly more associated with the 20th century, and second, other topics were slightly more prevalent in the 21st century. Lastly, the third group of topics seemed to recur more or less steadily throughout the whole period. In other words, on the level of topics in the abstract corpus, there seems to be no evidence pointing in the direction that sociological knowledge has become more disintegrated from 1980 to 2019. Rather, there seems to be a ‘core’ of topics that are repeatedly picked up by dissertations in sociology over time, which do not resonate with the idea of fragmentation.

111 Research question one reads: If fragmentation is understood as a historical process where the content of the knowledge produced in a discipline becomes increasingly disintegrated, do the corpora offer evidence that sociological knowledge is in a state of fragmentation?
Turning to the second research question,\textsuperscript{112} seeking to unravel whether there are paradigms in sociology, it would be faulty to conclude that all topics are seen as reflections of one or a few of the sociological dichotomies explained in the theory chapter. Indeed, about a quarter of the topics were interpreted to not pick up any words related to a dichotomy. In contrast, a larger number of topics seemed to resonate with one or a few poles of the dichotomies that this dissertation is utilizing to make sense of the results of computational text analyses of the abstract corpus so far analyzed in these chapters. Lastly, the remaining topics were understood to lie somewhere in between these two alternatives.

Recalling the tendency found in the previous chapter regarding the two communities of word constellations, an akin image of a dual-paradigm science emerged when investigating the relations between topics as expressed in the topic correlation network. On one side, we found all topics construed to be attuned to the qualitative community, and on the other, those considered to be related to the quantitative community. Further, the topics that were seen not to be connected to any dichotomy often expressed a strong correlation with one of the most representative topics for each respective community. These findings led to the interpretation that the same groups of abstracts that address these more empirical topics tend to also pick up words associated with topics adhering to one of the communities. Thus, the two communities of correlating topics seem to be associated with either the qualitative community or the quantitative community. This result appears to echo that of the previous chapter, and sustain the idea of a methodological divide in sociology discussed there.

Further, the results can be seen through the lens of the third research question,\textsuperscript{113} focusing on the social conditioning of knowledge. First off, as suggested in one of the previous paragraphs discussing fragmentation, the topics modeled on the abstract corpus can be said to be mildly affected by time. However, it was suggested that two-thirds of the topics expressed a rather clear association with specific departments and were therefore conditioned by place. There was one group of topics where abstracts from either Stockholm University or Umeå University scored the highest, and another where Uppsala University, Lund University, or Gothenburg University had the greatest effect. Lastly, the analysis suggested a group of topics that either

\textsuperscript{112} Research question two reads: Given that a paradigm denotes a crude replica constituted by reiterations of the same crude analogies in a discipline, do the corpora offer evidence that there are one, two, multiple, or no paradigms in sociology?

\textsuperscript{113} Research question three reads: Following the notion that the social conditioning of knowledge is expressed in the time and place of its production, do the corpora offer evidence that sociological knowledge is socially conditioned?
seemed rather unaffected by the aspect of place, or diverged in some aspect from the overall pattern.

When putting the tentative answers to three research questions together, we found that topics from both communities had a similar prevalence throughout the period with, perhaps, a small increase for the *qualitative community* during the 20th century and the *quantitative community* during the 21st century. For place, however, the topics correlating with the corpora of abstracts adhering to *Stockholm* and/or *Umeå* were all drawn from the *quantitative community*, whereas the ones coupled with *Uppsala, Lund, or Gothenburg* adhered to the *qualitative community* (with the exception of one isolated topic). Out of these two groups of topics, the ones taken from abstracts for dissertations defended in *Stockholm* and/or *Umeå* and positioned within the *quantitative community* appear to have a stronger effect on the topic distribution. Finally, the topics that were found to be more ‘neutral’ in terms of place were rather well-distributed between the two communities. Thus, this analysis of the effects of the dissertations’ defense year and defense location on the distribution of the 20 topics of this abstract corpus point out that the so-called methodological divide seems to be rather consistent over time, and conditioned on place.

Translated into the language of paradigm theory, we still find ourselves with the image of sociology as a *dual-paradigm science*. On this basis, one can perhaps speculate whether the crisis of sociology is less a reflection of the fragmentation of its knowledge production than of two opposing camps, diverging mainly on issues related to methodology. However, we have so far only looked at abstracts, which is the standard practice of computational studies of sociological texts presented in the chapter on the literature review. This is why the next two analytical chapters will be based on computations performed on full texts of sociology dissertations to investigate whether they express patterns that support or deviate from the findings and narrative presented so far in this dissertation.
The two previous empirical chapters were dedicated to answering this dissertation’s three research questions by analyzing a set of 815 sociology dissertation abstracts. Worth mentioning is that at this stage of the analysis (i.e., based on the tentative results presented in Chapter 6 and Chapter 7), we find ourselves with a depiction of sociology as a dual-paradigm science, characterized by a methodological divide between qualitative and quantitative studies. Each side of the divide seems to be intertwined with its selection of overarching word constellations and thematic arrangements that are conditioned by the place of sociological knowledge production. Further, a ‘core’ of word constellations and topics appeared to be rather consistent over time, which suggests that the prevailing understanding that sociological knowledge is fragmented might not hold true.

The present chapter will continue the investigation of these research questions. As was the case in the previous chapters, the analyses of paradigms presented here are interpreted through the lens that the dichotomies used in the theory chapter offer us. Further, regarding the social conditioning of sociological knowledge, time is here represented by the year the dissertation was defended and place by the university where the PhD defense took place. However, the analyses presented in this chapter will engage with the research questions using other methods and will do this with another corpus. The focus is on identifying the main styles of writing sociological knowledge in the full-text corpus, comprised of 380 sociology dissertations defended between the years 2000 and 2019. The previous chapters were based on an abstract corpus, as is customary in computational text analyses of scientific publications. In contrast to abstracts, the “full-texts” – the term used in academic literature to refer to the whole published text (with all its arguments, contradictions, and imperfections) – is an unaltered textual representation of, in this case, the dissertations, which in this corpus can be as long as 528 pages or 200,000 words (see Appendix B). However, in this and the next chapter, only the main bulk of the dissertations are subjected to analysis, since parts like reference lists can generate skewed and unwanted results.114

---

114 The main bulk of the text includes that from the first chapter to the last and, thus, excludes sections such as the foreword, table of contents, reference list, and appendices.
The length of the full-texts is a double-edged sword since the vast amount of text, on the one hand, opens up a new land of opportunities for finding hidden patterns in the corpus but, on the other hand, can be hard to boil down to something meaningful. Indeed, if the researcher tried, for instance, to map out the most common themes in such a corpus, the model would quickly be flooded by a wide diversity of common words that hold little meaning for the analysis. To recall, this was the main reason why the most common terms in language, the so-called “stop words”, were removed from the abstracts in the previous chapter to compute the structural topic model. However, some methods do not seek to escape the issue of commonality by excluding words but rather let them serve as data for mapping out relations between documents. At the center of this chapter, we find one such category of methods called computational stylistics or simply stylometry, which is frequently applied within the fields of computational linguistics and digital humanities (cf. Eder, 2017). One of the principal ideas of stylometry picked up in this analytical chapter is that authors are not distinguished by their usage of specific words but rather by their habit of repeatably writing a similar variety of common words when producing their texts (cf. Burrows, 2002). Thus, the leading notion is that by comparing the most common words of all full-texts of dissertations in sociology defended in Sweden, we can distinguish if the dissertations can be sorted into distinct styles of writing sociological knowledge.

Before we can present results of the analyses performed for this chapter, it is important to note the meaning of ‘style’ as it is used within stylometry and this dissertation. Style is here neither about the ‘literary style’ studied by literary critics, which often focuses on the effect the text has on its readers, intended or otherwise, nor about the normative version of the word as in good style or bad style. Style in stylometry is about word choices that are made repeatedly and which deviate from the ‘norm’ of the corpus. However, these are most often not conscious choices (e.g., to express individuality) but for some reason predetermined – like the historical context, the writing habits of authors, or that some phenomena tend to be expressed with certain types of words. Thus, for pedagogical reasons, ‘style’ will most often be referred to as ‘linguistic style’ in this dissertation to avoid confusion. This type of style or linguistic style can be defined as a feature of linguistic variance or, in even more simplified terms, recurring word usage (cf. Mairesse and Walker, 2011). Thus, the style also differs from the theme or content of a publication, which is often reflected in the keywords and labels attached to it. For instance, two dissertations situated within the sociology of work can write about the same topic but diverge in their linguistic style, and two dissertations writing about vastly different topics can resonate with the same overarching linguistic style (as expressed in their corpus). Worth mentioning is that the linguistic styles that will be mapped out in the corpus are under-
stood to be contingent on the corpus in the sense that they are generated from comparisons of words occurring in this specific collection of dissertations in sociology and, thus, the styles emerge inductively with no a priori qualities. See Appendix D for the technical details of the stylometric methods utilized in this chapter.

Going back to the research questions, the first one that this dissertation will answer in due course, focusing on the potential fragmentation of sociological knowledge, is investigated by exploring whether any distinct linguistic styles can be mapped out in the full-text corpus or not. The state of no recurring styles of writing sociological knowledge might in this case be interpreted as a sign of fragmentation. In contrast, the repetition of a linguistic style that can be interpreted along the lines of a few poles of the prevailing dichotomies of sociology would instead indicate the existence of a paradigm, which is the problem explored by the second research question. Thereafter, temporal and spatial conditionings of sociological knowledge, which is the focus of the third research question, are considered by exploring the distribution of the found linguistic patterns across time and place. Further, by leveraging the results of the previous analyses of word constellations and thematic arrangements in the abstract corpus, it is of interest to investigate whether the suggested divide between a qualitative community and a quantitative community is articulated in the full-text corpus.

The chapter is divided into five sections. First, the method of bootstrapped consensus trees is presented, which will be applied to model linguistic styles on the full-text corpus in the form of a stylometric network. Building on this analysis, the second section seeks to categorize the linguistic styles based on the most specific of their frequently used words vis-à-vis the other styles (z-scores). This analysis is expanded upon in the third section by close reading and hand-coding the most typical dissertations for each linguistic style based on the sociological dichotomies. The fourth section is a continuation of the previous three sections and investigates how the stylistic patterns mapped out are distributed over time and place. Lastly, section five is a discussion of

---

115 The first research question reads: If fragmentation is understood as a historical process where the content of the knowledge produced in a discipline becomes increasingly disintegrated, do the corpora offer evidence that sociological knowledge is in a state of fragmentation?

116 Such a style would, for instance, return to certain phrasings that include some types of phenomena and exclude others.

117 The second research question reads: Given that a paradigm denotes a crude replica constituted by reiterations of the same crude analogies in a discipline, do the corpora offer evidence that there are one, two, multiple, or no paradigms in sociology?

118 The third research question reads: Following the notion that the social conditioning of knowledge is expressed in the time and place of its production, do the corpora offer evidence that sociological knowledge is socially conditioned?
the implications of the analyses presented throughout the chapter for the answers that can be tentatively formulated to the three research questions that this dissertation has posed.

Discovering Styles with Bootstrapped Consensus Trees

As was mentioned in the introduction to this chapter, research that utilizes computational power to identify genres and perform author recognition proposes that linguistic styles are not found by looking into words most exclusive or specific to a text. The latter would often entail removing stop words\textsuperscript{119} that would cause ‘noise’ when attempting to single out the words that best characterize the theme being addressed by a collection of documents. In stark contrast, stylometry is based on comparing the most frequently used words in the corpus (cf. Eder et al., 2016), which to a great extent are comprised of stop words. This reasoning is derived from “a difficulty encountered by everyone who works in computational stylistics”, namely “that authors work at times in very uncharacteristic literary genres” (Burrows, 2002, p. 279). Based on research in the topic, an observation was made that instead of relying on a small set of variables that include only the most specific words used by a group of authors, “large sets of variables usually yield the most accurate stylistic signatures” (Burrows, 2002, p. 277).

Put into the context of analyzing stylistic patterns in a corpus of multiple authors, the hypothesis would be formulated along the lines that texts that share similar word frequencies are stylistically more similar or, put differently, express the same type of ‘linguistic style’. The idea of this chapter is to apply stylometric algorithms to the most frequently used words in the dissertations to model the strongest stylistic patterns running through the corpus in search of linguistic styles. The hope driving this analysis has been that such linguistic styles can provide insights into how sociological knowledge is patterned.

\textsuperscript{119} i.e., frequently used words that usually have little significance in natural language processing (e.g., “it”, “then”, and “could”) and, thus, are removed before training the model.
Figure 18. Stylometric network of sociology based on bootstrapped consensus trees, where each node represents a dissertation full-text, the node size reflects betweenness centrality, and communities are colored by modularity. The thickness of the edges connecting the nodes indicates how similar they are in terms of linguistic style.
In this section, bootstrapped consensus trees (BCT) are generated on the most frequently used words in the corpus to compute linguistic styles (for technical details, see Appendix D). The BCT – a method originating from research on language evolution and genetics (e.g., Rybicki and Heydel, 2013; van Dalen-Oskam, 2014; Stover et al., 2016) – in stylometry allows the researcher to run several intervals of different word frequencies in a single computation, where the process of bootstrapping\footnote{Bootstrapping is a method for resampling data by replacing them with random samples of the original data in order to measure the quality of inference (cf. Newitt and Hancock, 2001).} generates a set of virtual cluster analyses that are aggregated into a single consensus tree\footnote{A consensus tree is a graphical summarization of a set of trees that tries to represent the most frequently occurring characteristics for this set, which is the result of a computation involving a comparison of multiple clusters of trees (cf. Steel et al. 2000).} (cf. Eder et al., 2016). The benefit of BCT – in comparison to, for instance, a single principal component analysis – is that instead of presenting the result of a single run – which, for instance, opens up the problem of cherry-picking following the researcher’s hypotheses (e.g., Rudman, 2003) – several runs are performed and a consensus between these results are reached by singling out the most pervading characteristics. The idea, then, is that the best representation of the linguistic styles in the corpus is found in the most robust patterns of similarities between groups of documents, that is, those that tend to recur in multiple trees in contrast to more accidental similarities that tend to appear in only individual trees (see Appendix D for a more technical description of this method). The BCT computed on this full-text corpus is visualized in the form of a stylometric network where the nodes represent the dissertations and the edges show how similar texts are in terms of their word usages.

Figure 18 displays a stylometric network that is based on 31 runs of bootstrapped consensus trees covering the interval of the 50-1,550 most frequent words (with sequences of 50) in the corpus. The network is constituted of 380 nodes and the bootstrap consensus network algorithm generated 4,292 edges.\footnote{Only about one-fifth of the edges, those with the highest weight, are displayed in the network.} The average degree of the network (i.e., the mean number of edges per node) is 22.6, with the lowest value being 5 and the highest 126. In the graph, node size reflects betweenness centrality, where the bigger the size of the node the larger the role a full-text plays in bridging other full-texts in the network. In other words, a small node with a low betweenness centrality represents a more isolated text in terms of having a vocabulary consisting of words that are seldom shared with other texts in the network. The modularity algorithm had a score of 0.6 and generated five distinct communities, which are colored orange, purple, green, red, and blue in Figure 18.

\footnotesize
\[120\] Bootstrapping is a method for resampling data by replacing them with random samples of the original data in order to measure the quality of inference (cf. Newitt and Hancock, 2001).
\[121\] A consensus tree is a graphical summarization of a set of trees that tries to represent the most frequently occurring characteristics for this set, which is the result of a computation involving a comparison of multiple clusters of trees (cf. Steel et al. 2000).
\]
The most popular style in the graph is a *green community*, situated in the middle of the network that is made up of 130 nodes, which is more than a third of the total sample (n=380). Its position in the center, its large share of sizeable nodes (i.e., with a high betweenness centrality), and its interconnectivity – both within its community with the four other communities – suggest that this style might be relatively generic in its word usage. The second most populated community is constituted by 98 blue nodes (i.e., more than a quarter of all dissertations) and can be found at the right end of the network. The remote position of this *blue community*, the fairly equal size and relatively low betweenness centrality of its densely connected nodes, and the fact that it almost exclusively shares edges with two other communities give the impression of a more exclusive style vis-à-vis the *green community*.

On the other side of Figure 18, we find an equally sizeable community in orange, made up of 84 nodes. As was the case with the *blue community*, this *orange community* seems less central than the *green community* due to its position in the corner of the network, the relatively low betweenness centrality of its nodes (with a few exceptions), and its links with primarily green and blue nodes. For the latter, it is worth noting that the *orange community*, in contrast to the *blue community*, also shares edges with the *purple community*. A comparatively small community, made up of 54 red nodes (i.e., about one-seventh of all dissertations), is positioned in the low-middle parts of the network. The nodes of the *red community* are interconnected throughout the graph with the *green community* and take on the form of being separated into two main clusters. The cluster to the left has edges linking it with the *orange community* and, to a much lesser extent, to the *purple community*. In contrast, the cluster to the right shares edges with the *blue community*. The two blue clusters are linked with a handful of blue nodes, with edges spanning to the *green community*.

The abovementioned aspects of the *red community* suggest that it might be a less distinct community compared to the *orange community* and the *blue community*, but the reason why is not that straightforward. On the one hand, it is placed in the center of the network with two clusters of nodes, each interlinked with different communities, which might mean that it is less specific. On the other hand, it is constituted by relatively small nodes with low betweenness centrality in comparison to the green, orange, and blue communities, entailing that the community is not particularly interconnected with the other communities and thus more specific. For pedagogical reasons, the latter can be elaborated on with an example. For orange nodes, positioned at the left of the graph, to reach the blue nodes, positioned at the right of the graph, they tend to pass through green nodes rather than red nodes. Since the network is a representation of style, calculated based on how often authors
use different types of words, a simplified inference would be that, in general, the orange community and the blue community are stylistically closer to the green community style than the red community. In this context, the divided structure of the red community into two main clusters entails that one group of these dissertations is closer in style to the orange community and the other to the blue community. Thus, it might be that the red community is about as specific as the two other communities but is comparatively less stylistically unified. To investigate the linguistic styles further, we need to look at the words constituting each style to see what unites them and what separates them from each other.

Lastly, in stark contrast to other communities, we have a tiny and dense gathering of 14 purple nodes (i.e., almost only a tenth of the green community) positioned to the left in Figure 18. This purple community is constituted by relatively small and almost equally-sized nodes that sit in a distinct corner of the network. Their edges are primarily shared by the green community and, thereafter, the red community. The density, the position, and the low betweenness centrality of nodes of this purple community insinuate that it is the most peculiar of the five styles. Following this logic of interpretation, we would find the community in green at the other end of this peculiarity spectrum due to its central position in the network and its large share of nodes with high betweenness centrality. In between these two styles, we could position the blue community and the orange community since they are both stretching out towards the end of the network and have only a handful of nodes with a relatively high betweenness centrality.

To wrap up, this section initiated an investigation of how a computational text analysis of dissertations in their full-text format can be employed to uncover different styles of inscribing sociological knowledge. By adding modularity to a stylometric network based on bootstrapped consensus trees, five communities interpreted to represent five styles were generated. Through the position of each community within the network and the betweenness centrality of their nodes, some tentative differences were laid out. It was suggested that the green community, found in the very center of the network, is the most generic in this full-text dissertation corpus and the purple community is the most specific. The network is further structured by two relatively explicit styles, each on its diametral side of the network: the orange community and the blue community. The last style, captured in the red community, sits in the center of the network and is much less interconnected to the other styles in comparison to the green community, with one cluster linked to the orange community and another to the blue community. With this knowledge to hand, we can proceed with comparing the content of each style by looking into their most prevalent words, which was the basis for computing the network presented in this section.
Labeling Styles by Contrasting Z-scores

To obtain an impression of how to characterize the content of the five communities presented in the analysis above, we will now look at the most specific of the most frequently used words in each community of full-text dissertations. To recall, the communities were mapped out in a stylometric network generated by bootstrapped consensus trees computed on the most frequent words used in the corpus. This section is dedicated to exploring the most frequently occurring words preferred by each style through performing a contrastive analysis. Utilizing yet another tool from stylometry (cf. Eder et al., 2016), this analysis compares the z-scores of each community’s full-texts in a binary fashion to model their 20 most preferred words (see Appendix D for technical details). For the sake of accessible interpretability, the result is visualized through a single network in which the communities have been assigned a name reflecting their linguistic style (the orientation of their preferred most frequently used words).

In Figure 19, we find a network generated by 100 of the most frequently used words in the corpus, 20 for each community revealed by the bootstrapped consensus network, where the color of the edges represents the community preferring the words, and the thickness indicates the strength of the z-scores (in simplified terms how characteristic the words are for the community). Initially, we can see that the five communities occupy a distinct position in the network. To the northeast, we find the purple community, to the northwest the green community, to the west the orange community, to the east the red community, and to the southwest the blue community. We can see that some words are only connected to one community and others are shared by two or sometimes three communities. In this sense, the most distinct communities would appear to be the blue community and the purple community since they have the largest proportion of individual nodes, while the other three share nodes with all other communities, with the red community seemingly being the most integrated. To some extent, this pattern reflects the results unraveled in the analysis of the bootstrapped consensus network presented in the previous section. There, the blue community occupied one side of the network and both the orange community and the purple community the opposite side. In contrast, the red community and the green community were found at the center and most entangled with the other communities. So, let us look into whether this pattern is reflected in the words displayed in Figure 19.
Figure 19. A visualization of a contrastive analysis comparing the 20 most preferred words of the linguistic style sub-corpora (binary comparisons of z-scores) in the form of a network. The large nodes represent the linguistic styles, the small nodes words, and the edges z-scores (Craig’s extension of Burrows’ Zeta) colored by the style preferring the word.

Beginning with the words found furthest to the west of Figure 19, which denote the words that the 84 sociology dissertations classified as the style of the orange community tend to use, they are centered around the word PERSON. This word represents the label thought to capture the overall stylistic tendencies of this community, which is a person-oriented style, meaning that the dissertations tend to return to words that refer to the lifeworld of persons or individuals. To illustrate the reason behind this label, we can look at two categories of words found in the figure. The first category collects words interpreted as denoting human subjects in singulars – me, you, someone,
person — and plurals — children, people's — with a preference for female-coded words — her, she, woman. The second group of terms seems to be related to the very being of a person, namely as possessing — my, woman's — experiencing — I'm, feel, think, want, like, know, care — and performing actions — say, says, describe, mean. If we were to employ the theoretical lens offered in the theory chapter, the frequent words used by this group of dissertations seem to echo a subjectivist epistemology (objectivism/subjectivism) and a focus on human action (structure/agency). By taking into consideration the words interview, interviews, and interviewees, one could perhaps speculate that these dissertations tend to reference words associated with empirical studies (empirical/theoretical), qualitative methodology (quantitative/qualitative), and the micro level of society (macro/micro). Hence, the orange community is labeled — for the sake of this stylometric analysis — person-oriented style since it seems to primarily encapsulate dissertations addressing the narratives and experiences of human subjects.

Moving on to the far south of Figure 19, the most frequent words used by the blue community are found. The community has been named DISTRIBUTION, short for a distribution-oriented style, since these 98 dissertations appear to often use words that refer to quantities, varieties, and numbers. This tendency is explicitly found in a set of words interpreted to be used for measuring and comparing distributions — measure, measures, low, lower, high, higher, increase, average, rates, relative. Another tendency found here is that these dissertations tend to use words referring to individuals in the plural — individuals, women, family, household, population, countries. With the risk of over-interpreting the results, these two tendencies might, in stark contrast to the dissertations with a person-oriented style, signify a focus on social objects rather than social subjects (objectivism/subjectivism). This position is sustained by another category of words that can be found within statistics — data, effect, effects, factors, level, levels, variables — which further evokes a (natural) scientific vocabulary (science/humanities), quantitative methodology (quantitative/qualitative), an explanatory approach (explanatory/interpretative), and, perhaps, the presentation of empirical results (empirical/theoretical). Based on these collections of frequently used words, the blue community is called a distribution-oriented style. When juxtaposed against each other, it appears as dissertations with a person-oriented style tend to write about subjective experiences of being human, supposedly studied with a qualitative methodology, and dissertations with a distribution-oriented style address, to a greater extent, human individuals as objects in a dataset with, as the words entail, a quantitative methodology.
On the opposite side of Figure 19, the northeast corner is occupied by words used by the 14 dissertations in the purple community centered around the label concept, referencing that they express a concept-oriented style. Out of the words related to this linguistic style, the majority are interpreted as belonging to two potentially interrelated categories. On the one hand, there are words addressing the human consciousness or mental activity—consciousness, understand, thought, meaning, conception—which seem to be terms for expressing thoughts in human language—writes, critique—as well as male-coded words—he, him, his. Further, there are words more clearly related to scholarly work where some are more related to the humanities than others—history, sociology, philosophy, philosophical, theory, truth, reality, self. In vast contrast to the words preferred by dissertations applying a distribution-oriented style, the expressions found here do not seem to address empirical concerns but rather theoretical issues (empirical/theoretical). To speculate a little further, the words associated with this community appear to denote an interpretative methodology rather than an explanatory one (explanatory/interpretative), focusing on meaning and understanding, as well as the humanities rather than the (natural) sciences (science/humanities), with concrete references to philosophy and history. With this backdrop—and adding the occurrence of the words sociology, theory, and reality—the traditionally scientifically-coded terms (i.e., science, scientific, and object) could be interpreted as echoing a form of theoretical work in sociology. As was presented in the theory chapter, these types of studies tend to discuss the discipline’s relation to science and conception of reality. However, whether this is the case or not would have to be investigated further with other analyses, which will be conducted in the next section. Thus, the label concept-oriented style ought to be perceived as tentative at this stage of the analysis.

Heading over to the upper west side of Figure 19, we find the green community and words linked to the label relation, referencing dissertations with what has been labelled a relation-oriented style. This linguistic style encapsulates 130 full-texts with an over-representation of words referencing various constellations of relations. These can appear in a generic sense—system, organization, organizations, group, groups, people’s—which, in light of the fact that these are dissertations in sociology, probably denotes social constellations. Some indication of this interpretation can be found in the high prevalence of akin words that specifically refer to political arrangements—state, government, politics, sweden—where some indicate a focus on the distribution of power in society. In the light of the sociological dichotomies, these categories of words seem to denote social structures (structure/agency) transgressing the individual level (macro/micro). Further, the word relation is used to circle more specific terms for constellations of utterances and ideas—discourse, debate, perspec-
tive, knowledge, expression — as well as events and doings — practice, activities, action. In comparison to dissertations with a person-oriented style, the focus appears to be on the utterances and doings of not a single individual but rather a collective of individuals (i.e., supra-individual). Through this lens, the words shared with the concept-oriented style — e.g., science and scientific — are here interpreted as other examples of words referencing collective endeavors founded on social relations that might be speculated to lean towards, or related to, science (science/humanities).

Lastly, the red community can be found to the east of Figure 19, centered around the label CIRCUMSTANCE, capturing 54 dissertations with a circumstance-oriented style. At the center of this linguistic style lies words interpreted mainly to reference the condition or context of phenomena. First, there are general words that circle the most fundamental forms of circumstance like temporality — day, years, period — and spatiality — local, space, spaces. The reading here put forth is that of an over-representation of nouns and adjectives used for characterizing or situating the study object within a context of interest. Examples include the usage of words referring to different dimensions of society — community, city, country — including the political sphere — policy, policies, rights, political, politics, government — and the economic context — economic, market. We further find some words with cultural connotations — ethnic, ethnicity, foreign — that tend to be used to emphasize how people are conditioned. Further, there are indications of spatially and juridically bounded spaces and identities — national, international — where some words are more empirically specific — sweden, swedish, european. This includes examples of even more particular circumstances that might condition the study object like war or migration. In addition, there are a few words — interview, interviews, interviewed — suggesting that some of the dissertations with a circumstance-oriented style engage in empirical studies (empirical/theoretical) conducted with qualitative methods (quantitative/qualitative).

Although this contrasting analysis of the 100 most frequent words cannot capture all nuances of the stylometric network, it has managed to find characteristic tendencies to identify some of the main differences between the five linguistic styles. By characterizing how the most frequently used words in the corpus differed from one community of dissertations to the others, the five communities mapped out in the stylometric network became coupled with an informative label to signify the frequent words typically used. While the orange community is titled person-oriented style since it was constituted by words referring to human subjects with their experiences and doings, the blue community is given the label distribution-oriented style for primarily gathering words for comparing quantities, populations of people, and stas-
tical methodology. In contrast, the smallest of the communities, the purple community, is termed concept-oriented style and found in a distinct position in the company of words for concepts expressing thoughts and interpretations, commonly found within several branches of the humanities, like philosophy and history. Further, the green community is called relation-oriented style for mainly addressing words related to social constellations of people, practices, and utterances. Lastly, the red community is referred to as circumstance-oriented style because it was shown that the dissertations in that community tend to use words associated with the condition or context of social phenomena.

At this stage of the analysis, perhaps the best case for a stylometric representation of a methodological divide seems to be expressed in the distribution-oriented style and the person-oriented style, with the former playing the role of the quantitative community and the latter that of the qualitative community. This is because they are both directly referencing words for empirical studies, and each community is associated with opposing poles in the dichotomies of quantitative/qualitative, objectivism/subjectivism, and macro/micro. However, these two communities only make up less than a third of all dissertations in the full-text corpus.

If taking into consideration the three other styles – the concept-oriented style, the relation-oriented style, and the circumstance-oriented style – a more complex pattern emerges. For instance, the circumstance-oriented style is as associated with words for qualitative methods as the person-oriented style, yet cannot be said to be associated with the micro level or subjectivism. Further, the concept-oriented style appears to be the community of dissertations most strongly associated with words related to theory, interpretive methodology, and the humanities, yet does not seem to be referencing words related to qualitative methods or the micro level of society. Similarly, dissertations categorized as belonging to the relation-oriented style appear to be about social structures and the macro level of society but express no clear methodological tendency. Thus, these three styles are harder to force into a single master dichotomy. Indeed, the pattern found when looking at all five dichotomies transpired to be more intriguing than expected, and included more dimensions than a clear-cut division between a qualitative community and a quantitative community. The differentiations and similarities between the linguistic styles are explored even further in the next section, which is centered on a close reading of the most typical works for each style; a reading performed because the limits of computational text analyses may have been reached at this point.
Differentiating Styles with Close Reading

In this section, the five dissertations that are most typical for each style are identified and examined to deepen the understanding of how they differ from each other, as proposed in the previous section. Within the context of this analysis of styles in sociology, the term “typical” refers to dissertations that express characteristics that are specific to a style. This is measured through the stylometric network visualized in Figure 18 by singling out the 25 dissertations with the highest degree within its assigned community and the lowest degree outside that community. The intuition behind this choice is that these dissertations are supposedly the ones that most exclusively share the specific frequently used words characterizing a style. Thus, for each style, five full-texts were analyzed through close reading and hand-coding. The close reading focused on the aims, research questions, data, methods, and themes of the full-texts of sociology dissertations, and were coded based on the sociological dichotomies presented in the theory chapter. The codes were then compared to each other to find the most defining characteristic of each style in comparison to the others.

As was addressed in the literature review, this has been one of the most popular methods for sorting out patterns of knowledge from publications in sociology (e.g., Erola et al., 2014; Payne et al., 2004; Platt, 2008; Schrecker, 2008). One of the reasons for adding this method to the analyses performed so far is to add human-based validity to the distant reading tasks performed by statistical and machine learning models thus far (cf. Moretti, 2000). The sub-study presented in this section follows a simple structure where the most distinctive features for each style are spelled out one by one, in the same order as in the section above. A conceptual map for simplifying the analysis put forth will be given at the end of the section.

In the previous section of this chapter, we found that the person-oriented style was constituted by words referring to human subjects with their experiences and doings, as well as interviews. When reading the five most typical dissertations in this style, it is clear that they all share an obvious empirical focus (empirical/theoretical). Indeed, the research questions presented in the dissertations are centered on human subjects and their concrete actions being observed rather than any form of armchair sociology with abstract theoretical aims, and the sociological classics are scarcely mentioned. All studies in the typical dissertations for the person-oriented style apply qualitative methods (quantitative/qualitative) that are applied to study micro-social phenomena (macro/micro) as they unfold in situ. Methods are primarily formal interviews, but some studies also include participant observations, covering a partaking in various activities performed by the informants as well as infor-
mal talks. In one case, the informants are even described to be factual co-researchers who conduct individual data collection. Thus, the person-oriented style seems to be applying subjectivist epistemology (objectivism/subjectivism) since it appears to deal directly with developing knowledge of the concrete experiences of the body and meaning-making of the mind, as they are expressed by living agents (structure/agency). The shared thematic thread of this sample can be labelled as social problem research that often includes different types of value judgments (fact/value). This includes questions about how identities such as gender roles, ethnic categories, disabilities, or victimhood are experienced and negotiated through social interaction.

Turning to the distribution-oriented style, it was suggested that the linguistic style gathers words for comparing quantities, populations of people, and statistical methods, and the key works do indeed stand out to be representative of the other styles in a few related aspects. First, the dissertations that have been read and hand-coded for this analysis are all comprised of four independent empirical studies (empirical/theoretical). In contrast, the dissertations in the other four linguistic styles are comprised of one larger comprehensive study. Second, the main data in the typical dissertations for the distribution-oriented style are generated through immense surveys of tens of thousands of individuals, some of which are part of comparative studies of different countries, and leverage register data that cover more or less the complete national population (macro/micro). Thirdly, the methodology in these typical dissertations for this style is formulated and exclusively drawn from quantitative methods in general, and statistics in particular (quantitative/qualitative). Following the format of the data and the methods, individuals naturally become objects that fit into social categories rather than autonomous subjects full of contradictions (objectivism/subjectivism). Lastly, the studies are presented through a scientific vocabulary (science/humanities) and with an explicit intention to explain mechanisms rather than to interpret phenomena (explanatory/interpretative) – effects are measured, hypotheses are tested, and so on – that differs from dissertations in the other style. Therefore, not surprisingly, there are few contemplations on the nature of sociological knowledge or references to the classics of sociology in the typical dissertations in the distribution-oriented style that have been hand-coded in this analysis, and the studies conducted in these dissertations are conceptualized as adding knowledge to a specialist rather than a generalist research front. A recurring theme unifying these typical dissertations within the distribution-oriented style is social structures of inequality (structure/agency), particularly the relationship between individuals’ social position (class, gender, etc.) and their health or work outcomes (holism/individualism).
The concept-oriented style stood out in the preceding analysis since it was so clearly drawing on concepts found within the humanities. When exposed to guided close reading and hand-coding, the most striking aspect of this style, in comparison to all other styles, is its complete absence of empirical material in the social scientific sense of the word (empirical/theoretical). Another distinct feature of the typical dissertations in the concept-oriented style is the position of the sociological classics. These authors and their works are both more present in mere quantitative measurements and more central in qualitative terms, meaning that the main themes in the dissertations and, most often, their research questions involve concepts found in classical works within sociology. A third main characteristic of the concept-oriented style is its arrangement seemingly drawing on a humanistic tradition (science/humanities). The typical dissertations hand-coded in this analysis are all monographs based on formulating a narrative stretching over 200-300 pages, founded upon reading and discussing key works of authors. For one-half of the dissertations, the strategy is to construct a fundamental notion of sociological knowledge based on classical thinkers, carve out a problem in their conceptualization of sociology, and challenge it with a theoretician interpreted to be unfamiliar within the tradition.

The other half of these typical dissertations take a new perspective on the orthodox questions of sociology. In both cases, the form of close reading performed involves treating the texts not as fluent social constructs but as something closer to “containers of knowledge”. In this sense, the methodology is akin to traditional exegetics that study “the Word” rather than propose prophecies, but the religious fact is here replaced with a sociological fact (fact/value). Examples include what distinguishes the experience of modernity, and what is the relationship between sociological knowledge and the common sense of everyday life, drawing heavily on social psychology. For both halves, deep interpretations of literature (explanatory/interpretative) focus on the subjective reality (objectivism/subj ectivism) and individuality of humans (structure/agency). Echoing the position of classics like Weber and Durkheim, the conclusions presented in the full-texts are not characterized by value judgments of how society ought to be changed but by arguments for the disciplinary relevance of the interpretation of the factual books put forth (fact/value).

Recalling the foregoing analyses, the relation-oriented style was seen as the largest and most generic style in the full-text corpus that this chapter is based

---

123 Empirical material can here entail data found within an array of sources ranging from documents, field notes, and interview transcripts to register data and surveys.

124 Recurring names in these dissertations’ theoretical underpinnings are the international classics George Herbert Mead, Georg Simmel, and Max Weber as well as the Finnish-Swedish modern classic, Johan Asplund, to name but a few.
on. It picked up a wide range of terms related to arrangements of social relations spanning from the level of ideas or discourses to the level of practices or events. When reading and comparing the most typical dissertations in the relation-oriented style with those of the other styles, one of the most distinctive features is the deep presence of discourse analysis and the heavy theoretical vocabulary that comes with it. The empirical material constituting the discourses studied within these typical dissertations is almost exclusively constituted by texts, both in the form of traditional documents like news articles and books, as well as interview transcripts. These texts are not pre-processed and quantified (e.g., words are not converted into tokens to compute vectors as above) but rather scrutinized qualitatively in search of overarching patterns that can illuminate a larger phenomenon (quantitative/qualitative).

For about half of the typical dissertations that constitute the sub-sample for the analysis that this specific section focuses on, it is not explicitly stated in the full-text how this method is conducted, but one can expect the form of hermeneutical reasoning that is typical for discourse analysis. The other half of the dissertations describe the method as a process of close reading and hand-coding (i.e., much like the analysis presented in this section). In either case, the texts and actions of the study participants do not appear to be treated as mirrors of their subjective reality (objectivism/subjectivism). Yet, at the same time, the analyses study social constructs that reject the idea of an objective reality “out there” (objectivism/subjectivism). However, what is clear is that the focus is set to delineate structured positions in societal discourses and practices (structure/agency). The themes dealt with in these five dissertations include the politics of sciences and constructions of identities like gender, race, and age. Further, all the typical dissertations in the relation-oriented style have clear theoretical aims (empirical/theoretical) located on an abstract macro level of society (macro/micro). The aims are met by a critical analysis of how ideas and material phenomena are articulated, as well as how they ought to be conceptualized sociologically (according to the authors, that is). In the concluding sections, the dissertations presented more or less explicit value judgments derived from the results of the study that state how they want to “change” the social world in a “better” direction (fact/value).

Lastly, we have the circumstance-oriented style, which was shown to be a smaller linguistic style divided into two parts and interpreted to capture dissertations with a focus on the circumstances in which study objects and social phenomena are situated. These include the influence of different spheres of society on identities and societies, which include the temporal, spatial, political, economic, and cultural context. The close reading and hand-coding of the most typical dissertations of the circumstance-oriented style reveal
that all five are applying the methodology of ethnography. Indeed, the dissertations are based on fieldwork (that is more or less extensive and longitudinal) that is constituted by a combination of methods such as interviews, informal talks, and participatory observations in the respective field of study (quantitative/qualitative). In all cases, the qualitative approach is sustained by contextualizing the field of study with 'hard' data like descriptive statistics (quantitative/qualitative). The cases included in these dissertations are non-governmental organizations, educational institutions, international companies, or diaspora. Thus, the typical circumstance-oriented style dissertations seem to focus on the meso level (macro/meso/micro) by targeting social phenomena that are constituted by the actions of a group of individuals. The aims of these typical dissertations are formulated from a processual perspective, where the respective studies present "thick descriptions" (explanatory/interpretive) of how a group or an organization undergoes certain changes. This transformation is then explained by referencing external conditions (explanatory/interpretive). The overarching themes are social movements, education, work, migration, and commerce. While about half of the dissertations state an aim involving only the empirical realm (empirical/theoretical), the rest include a clear theoretical focus (empirical/theoretical). The dissertations are ethnographies of the meso level and involve understanding the ‘inner world’ of their interlocutors (structure/agency), but seem more drawn towards delineating the structural constraints of society than the concrete individual experiences studied (structure/agency).

The suggested relations between the linguistic styles and the sociological dichotomies can be simplified by the conceptual map found in Figure 20. In other words, this is a visualization of what the results of the deep reading and hand-coding of the typical dissertations for each of the linguistic styles identified through stylometry suggest in terms of the dichotomies that inform the interpretation of results in this dissertation. In the figure, the 25 typical dissertations are represented by their assigned linguistic style. For example, the label CONCEPT found in the northwest corner represents the five dissertations analyzed with a concept-oriented style, which used the method of close reading. The map is constituted by two primary axes that indicate the overall epistemological and ontological direction of the five linguistic styles. While the x-axis describes whether the dissertations are more focused on the agency of humans or the constraints of social structures (structure/agency), the y-axis describes whether the aim of the studies is mainly dedicated to sorting out an empirical or a theoretical problem (empirical/theoretical).
To recall, one of the most distinctive features found for dissertations with a *person-oriented style* was interview-based studies of the experiences and actions of humans. In contrast, the dissertations with a *concept-oriented style* were described to deal with theoretical problems, yet they share a focus on human agency rather than social structures. Thus, these two styles are both situated on the same side of the *structure-agency* dichotomy but at opposite sides of the *empirical/theoretical* dichotomy. To continue, the dissertations categorized as having a *distribution-oriented style* are interpreted to have an aim at the empirical level and study social structures through a quantitative methodology (statistical analysis). The dissertations with a *relation-oriented style*, on the other hand, share this focus on social constraints in the form of supra-individual discourses (discourse analysis) but are driven by a heavy theoretical aim. Lastly, the *circumstance-oriented style* was found to include representatives on both sides of both axes, and is therefore positioned in the middle of Figure 20. To expand with an example, the style is interpreted to be more focused on human action than the abstract discourses often targeted by the *relation-oriented style* (*structure-agency*) but more drawn towards
delineating the constraints of society than the concrete individual experiences studied within the person-oriented style (structure/agency).

Figure 20 also includes two other dichotomies that illustrate contrasting features between two linguistic styles based on the analysis of the most typical dissertations. Perhaps the most intense distinction in the corpus can be found between dissertations with a concept-oriented style versus those with a distribution-oriented style. The gist of this distinction can be found in the science/humanities dichotomy. Dissertations with a concept-oriented style were found to focus on abstract human constructs that are analyzed with a philosophical and historical method and, thus, echo disciplines traditionally seated within the humanities (science/humanities). On the other side of the dichotomy, we find dissertations with a distribution-oriented style since they articulated an epistemology and a methodology that mirror the (natural) sciences (science/humanities). Somewhere between these two ‘extremes’, one can identify the typical representations of a circumstance-oriented style. This is because these dissertations were led by a qualitative methodology that, on the one hand, had a clear scientific pretension but, on the other hand, found it preferable for the researchers to immerse themselves in the field and generate “deep interpretations”.

The other example of two seemingly antagonistic positions can be illustrated with the macro/micro dichotomy. To recall, the dissertations with a person-oriented style were interpreted to study the life worlds of ‘concrete’ individuals and are, thus, seen to be associated with the interactional or micro level of society (macro/micro). In comparison, dissertations with a relation-oriented style study more abstract, supra-individual entities like discourses and systems of practices, and are therefore interpreted to be associated with the macro level of society (macro/micro). The dissertations representing a circumstance-oriented style were found to focus on meso-level phenomena like communities and associations of people, and are, therefore, positioned in the middle of this dichotomy (macro/meso/micro).

Lastly, the figure comprises a two-sided discrepancy that is accentuated by different positions on the fact/value dichotomy and includes elements that might be more provocative for some readers. Starting with the least controversial point, dissertations with a distribution-oriented style – argued to be dedicated to empirical aims, social structure, and scientific rather than humanistic research – tend to concentrate on “hard facts” in the representations of tables and graphs when presenting their results and analyses (fact/value). In stark contrast, representatives of a person-oriented style were found to present studies that engage and often sympathize with the struggles of their study subjects. Thus, these types of analyses tended to be entangled with different forms of value judgments (fact/value). On a more abstract and im-
plicit level, the dissertations indicated to be the most typical for a relations-oriented style also tended to be value-driven (fact/value). The included studies are critical analyses of society that presented various postulates of how the social world should be changed to be enhanced, which intended some form of a fairer society. Further, dissertations with a concept-oriented style are also found to be close to the fact pole (fact/value), despite being interpreted as the diametric opposite of a distribution-oriented style on the four other dichotomies mentioned. This is, however, a different version of “facts” at play here. Retracting to the reading of the dichotomy put forth in the theory chapter, two of the main classics of sociology, Durkheim and Weber, argued for getting rid of value judgments in sociology (fact/value). Similarly, these dissertations with a concept-oriented style, building heavily on the legacy of the classics, seem to approach the books and theories they interpret not like arbitrary and fluent social constructs, but as seemingly factual “vessels of knowledge” frozen in the time and place of their creation (as is told by the current historiography, that is).

To conclude this section based on reading and hand-coding the five most typical dissertations for each linguistic style, it was suggested that the main commonalities and differences between the five styles can be simplified within a single representation found in Figure 20. In this figure, each style was given a position within an area of tension constituted by the two main axes, derived from one epistemological dichotomy (empirical/theoretical) and one ontological dichotomy (structure/agency). Here the dissertations with a person-oriented style or a concept-oriented style were found to more clearly express an ontological focus on the experiences and actions of human beings. In contrast, the constraints of supra-individual structures were interpreted to be at the heart of dissertations associated with a distribution-oriented style or a circumstance-oriented style. Further, the typical dissertations of the person-oriented style and the distribution-oriented style were said to share a clear emphasis on posing questions to concrete phenomena scrutinized by the tools of empirical research. In comparison, those associated with a concept-oriented style and a relation-oriented style primarily dealt with generating knowledge on theoretical entities. In addition, the dissertations with a circumstance-oriented style took up an intermediate position in the very middle of the area of tension. Lastly, the dissertations’ relation to three additional dichotomies – one ontological (macro/micro), one epistemological (fact/value), and one methodological (science/humanities) – were discussed to add further nuances to the distinctive features of the five linguistic styles.

With this categorization of the linguistic styles’ main currents in terms of content in place, we can move on to investigate the social conditioning of each style in terms of their prevalence in time and place. In this analysis, the
results generated in the previous analytical chapter on abstracts for dissertations in sociology are drawn upon to enhance the understanding of how the linguistic styles are distributed in the full-text corpus.

Spatial and Temporal Distributions of Styles\textsuperscript{125}

With the five linguistic styles and an interpretation of their general characteristics at hand, this section serves the purpose of investigating how the linguistic styles are distributed over time and place. For this task, we return to the two metadata used as points of departure in the previous empirical chapters – the year when the dissertations were defended and the universities where the defense took place – but now target the full-text of the dissertations instead of their abstracts. The technique applied in this chapter is based on comparing descriptive statistics of the metadata with a third one, the linguistic styles. To recall, the latter is based on the interpretation of the community each dissertation was allocated to within the stylometric network presented at the beginning of this chapter. In this analysis, the results of the previous chapter will be summoned to add additional context. Inevitably, this move raises the subordinated question of whether the patterns found in the abstract corpus for the 21\textsuperscript{st} century are reflected in the full-text corpus or if they tell incompatible stories. Thus, before diving into the analysis, the reader might benefit from a quick recollection of the results presented in the two chapters that preceded this one.

The general pattern of the dissertation abstract corpus appeared to be one of a division between a \textit{qualitative community} and a \textit{quantitative community}. This methodological divide seemed intertwined with a set of themes that were shown to be fairly consistent over time, with a few exceptions that were more temporally conditioned in the sense that these themes were more popular at the beginning (the 1980s) or the end (2010s) of the period studied or vice versa. However, when observing how the division and the thematical trends presented themselves over place, a clear separation was proposed.

\textsuperscript{125} In the previous chapters it was clarified that the variable of place has been added to the analyses because it is believed to have an impact on the kind of sociology that sociologists conduct in different parts of the world (e.g., Aaltojärvi et al., 2008; Bjarnason and Sifusdottir, 2002; Collyer, 2013a, 2013b, 2014; Moksony et al., 2014; Zougris, 2019). Thus, although spatiality is also presumably interesting for readers who are familiar with the national context from where the data come from (i.e., the universities in Sweden where the sociology dissertations analyzed stem from), this analysis is not about Swedish PhD programs in sociology per se but rather about the social conditioning of the sociological knowledge they generate in the form of dissertations. Thus, because space is hereby used as an angle for theorizing, rather than being primarily a backdrop to the results, information about the Swedish context is offered in Appendix A.
Indeed, it was suggested that the abstracts of dissertations from two universities were over-represented in the *quantitative community – Stockholm University and Umeå University* – and two in the *qualitative community – Lund University and Uppsala University*. In addition, abstracts from the last university – *Gothenburg University* – were interpreted to be positioned in the middle of the methodological divide. This section will be a continuation of the structure found in the previous chapters in the sense that we will first investigate the temporal distributions of the linguistic styles and thereafter head on to scrutinize the relationship between style and place.

Figure 21. Stacked area chart showing the temporal distribution of the five linguistic styles in absolute numbers. Each style is displayed in proportion to the gray area which depicts the total corpus and is constant in all five parts of the chart.

Figure 21 is made up of five stacked area charts that indicate how prevalent different styles have been over the period included in the full-text corpus, that is 2000-2019. Thus, it is based on corpora divided by linguistic style and sorted annually. Each chart in this figure displays the relationship between the yearly count of dissertations falling into one stylistic category (each style in its assigned color) and includes information on all dissertations defended in the same year (seen in gray). When comparing the linguistic styles to each other, it appears that dissertations conveying a *relation-oriented style* are the most popular throughout the period and are followed by those echoing a *person-oriented style* or a *distribution-oriented style*. Thereafter, there is, in figurative terms, a step down to dissertations applying a *circumstance-oriented style* and a dive down to the few examples resonating with the lin-
guistic traits of a concept-oriented style. This finding is expected considering the results generated in the previous analyses presented in this chapter. While all styles to some extent fluctuate over time, with some noticeable peaks and valleys, the prevalence of each style appears to be fairly constant over time. The major exception to this general trend is the concept-oriented style, which had a small presence in the 2000s, only to completely disappear in the early 2010s.

Looking back on the results presented in the previous empirical chapters, it was there proposed that time had a rather modest impact on the structure and content of the observed methodological divide that was mapped out in the abstract corpus. It was suggested that the overarching thematic arrangement of the corpus was one where the 20 topics generated by the structural topic model were entangled in the divide. While many themes appeared to be quite consistent over time, there were a few themes that expressed a trend that was ascending (discourse/politics, micro/humanist, health/demography, and family/background) or descending (history/economy, theory/knowledge, space/culture and, to a lesser extent, ethnic/migration and domestic/narrative) over time. For the former, we could see that the topics interpreted to be at the heart of the quantitative community – related to demography, health, and social categories of larger populations – increased in popularity during the 21st century. The same seemed to ring true for some topics deemed to be central for the qualitative community – primarily themes addressing politics, discourses, and the micro level. Thus, we would perhaps expect these thematic arrangements to be expressed on the level of the linguistic styles as well. In particular, one could imagine that the distribution-oriented style, and, to a lesser extent, the relation-oriented style and the person-oriented style, would be popular in the full-text corpus that this chapter analyses. In contrast, topics related to history, theory, knowledge, space, and culture seemed to express a downward trend from the 20th century. Thus, it is possible that the concept-oriented style and the circumstance-oriented style share a modest prevalence in this corpus. With this recollection at the back of our minds, let us now analyze the temporal manifestation of each style.

Let us return to where we left off in analyzing Figure 21, in the tiny concept-oriented style that is only present in the first half of the time period. Considering the decline of the thematic arrangement made up of words related to history, knowledge, and theoretical investigations in the abstract corpus, we are given some context as to why this style, which picks up a similar type of keywords (among other types of words), is so small in general and why it is not prevalent at the end of the 2010s. Not only does this further support the interpretation that the style is distinct but it also suggests that it seems to be a style of writing sociology that, figuratively speaking, belongs to the past rather than the future. It might be worth mentioning that different versions of
stylistic bootstrap consensus trees were run, and on top of these modularity algorithms with different presets, and in all cases, a distinct community constituted by the same small selection of dissertations that constitute the concept-oriented style was found.

Contrary to the concept-oriented style, the circumstance-oriented style in red was interpreted to be an unspecific style, which in Figure 21 seems to slightly increase in prevalence over time, expressing a higher presence in the 2010s than in the 2000s. In line with the general trend of the whole corpus, the style is seen to have its uttermost peak in 2011, but it also shows a notably high value in 2019. This pattern appears to be somewhat puzzling concerning the thematic arrangements found in the dissertation abstract corpus as presented in previous chapters, where the culture topic decreased over time, yet, at the same time, the topic constituted by words for ethnicity and migration slightly increased at the end of the 2010s. Nevertheless, one has to remember that full-texts include a much greater variety of terms than abstracts, and that styles are based on wide-ranging patterns in the most commonly used terms (i.e., not computed by comparing specific words like in the structural topic model based on the abstract corpus). For the circumstance-oriented style, these terms were interpreted to signify a somewhat betwixt position by addressing a wide range of tokens either associated with both poles of sociological dichotomies or that seemed to trespass these dichotomies. Therefore, it would make sense that the style follows the general pattern of the corpus.

Leaning on thematic arrangements exposed in the abstract corpus, we would perhaps expect to see a substantial occurrence for the distribution-oriented style, the person-oriented style, and the relation-oriented style in the full-text corpora analyzed here. This is indeed the case, which was also expected based on the results generated from previous stylometric analyses in this chapter. While the three styles share a consistent prevalence over the full period, the relation-oriented style in green, which is the largest style, shows a trend closest to the general trend of the corpus displayed in gray. Thus, since the style reappears to the same extent as in sociology dissertations in general, this result seems to reaffirm the interpretative position taken earlier in this chapter, holding that the style might be the most generic of the five in this sociology dissertation corpora. The person-oriented style exposes a different pattern with its highest scores in the middle of the time period, 2010, followed by 2014. In contrast, its lowest relative prevalence can be found at the beginning of the time period, before 2004, and even more so at the end, after 2016. It would be a stretch to suggest that these tendencies reflect a style in decline, like that of the concept-oriented style. Rather, it would be more accurate to propose that the person-oriented style, which is construed to be the best representative of a qualitative community because it picked up
words related to the experiences, actions, and narratives of human subjects as well as the method of interviews, is associated with the years 2008-2015.

Compared to the person-oriented style, which to some extent reflects the shape of an inverted U, the distribution-oriented style of sociology dissertations seems to be more U-shaped with its largest peaks in 2001 and 2017. Moreover, the style appears to express relatively high values over the whole time period, with a small tilt towards the right end. Of the two styles, the distribution-oriented style is the one diverging the most from the general trend of the corpus, which peaks at the break of the 2010s. This is notable since the distribution-oriented style is seen as the clearest representative of the quantitative community as it was shown to draw on a multitude of terms for comparing quantitates as well as scientific terminology, and even concrete words for quantitative methods, like surveys. When comparing the distribution-oriented style with the relation-oriented style, which is taken not to be as clearly associated with methodology but to lean towards the qualitative community due to its seemingly close relation to discourse analysis, we can see that this style peaks at the distribution-oriented style’s lowest points and drops at its highest points. This adds up to the impression that words related to the qualitative community and quantitative community are not equally distributed over this rather limited period.

These temporal patterns can be enhanced by moving beyond the idea of a methodological divide and utilizing the idea expressed in Figure 20 of two axes spanning from a theoretical focus to an empirical focus, and from a focus on the agency of humans to a focus on the structures of society. To recall, it was suggested that dissertations in the concept-oriented style and the relation-oriented style resonate with more theoretical sociology whilst the distribution-oriented style and the person-oriented style echo more empirical sociology. As for the other axis, dissertations in the person-oriented style and the concept-oriented style were interpreted to be more attuned to human agency, and the distribution-oriented style, as well as the relation-oriented style, were more attuned to social structures. The circumstance-oriented style here stands in the middle of both axes. By comparing the five styles through this lens, there is some evidence for the thesis that the 2000s was a period during which sociology dissertations in Sweden favored a style of writing that addressed sociological aspects of theory and human agency, while the 2010s seemed to draw more heavily on the empirical domain and to address social structures. With this general understanding of how the linguistic styles are distributed over the era covered in the full-text corpus, we can head on to investigate the relationship between the five styles and the five main universities.
Figure 22. Stacked bar chart showing the spatial distribution of the five linguistic styles (each university corpus adds up to 100%).

Figure 22 shows the distribution of full-texts over the variables of styles and place, in this case, the prevalence of each labeled style vis-à-vis each university. At first glance, we can see that no sociology dissertation style is solely associated with a single university but rather spread out over all five universities. Indeed, in most cases, the universities have at least a few dissertations in each style – the exception being the relatively small concept-oriented style, where neither Gothenburg nor Umeå dissertations are represented, and Uppsala and Lund where this style is most prevalent. Another noticeable spatial feature is that all five universities appear to show high scores in a combination of two of the three most prevalent styles in the full-text corpus (i.e., the distribution-oriented style, the person-oriented style, and the textual style). The combination of the two styles makes up 65-78% of each university’s total dissertation corpus and in all five cases, the relation-oriented style is one of the two most prevalent styles. For dissertations defended in Stockholm and Umeå, the role of the other top style seems to be filled by the distribution-oriented style, and for the other three universities – Gothenburg, Lund, and Uppsala – it appears to be the person-oriented style. Worth noticing, though, is that dissertations defended in Gothenburg, again seem to take a middle position since this university also has a sufficient number of dissertations in the distribution-oriented style. In addition, all universities have a considerable share of dissertations in the circumstance-oriented style, with
the largest group of dissertations in this style stemming from *Uppsala* and *Lund*.

Recalling the analysis put forth in the previous chapters, it was suggested that there might be a correlation between the spatial categories and the principal thematic arrangements of the dissertations. In simple terms, the analysis spelled out that the abstract corpus seemed to be split into two university groups – one filled by *Stockholm* and *Umeå*, the other by *Lund* and *Uppsala* – that seemed to be associated with the *quantitative community* and the *qualitative community*, respectively. By inspecting Figure 22, however, it appears to be the case that full-text dissertations from these universities are not as clearly defined by the same spatial separation in style as was expressed in the analyses of the abstract corpus on word constellations and thematic arrangements. Nevertheless, the stylometric results presented in this chapter seems to some extent to resonate with the methodological divide, which is worth discussing in more depth, style by style.

As was explained above, dissertations in the *distribution-oriented style* were understood to be characterized by empirical (*empirical/theoretical*) comparisons of quantities (*quantitative/qualitative*), particularly large populations of people (*macro/micro*), tend to focus on social structures (*structure/agency*), and draw heavily on a scientific vocabulary (*science/humanities*). Thus, these dissertations can be interpreted as being closely related to the *quantitative community* of the methodological divide. By comparing the distribution of this dissertation style over the universities, a line could be drawn between this full-text analysis and the results found in the analyses of the abstract corpus presented in the two previous chapters. Indeed, the dissertations dominating the *distribution-oriented style* come from *Stockholm* (51%) and *Umeå* (37%), while *Lund* (7%) and *Uppsala* (6%) contribute with the smallest number of dissertations in this style. In between the two university groups we find *Gothenburg* (18%), which was also the case in the analyses of the abstract corpus.

Following the discussion put forth in the previous section of this chapter, things do, however, get trickier if trying to single out a style that would represent the qualitative side of the methodological divide. To some extent, all of the four remaining styles share some connotations of qualitative sociology, yet one was interpreted to do so more exclusively: the *person-oriented style*. This sizeable style was understood to resonate with words related to the experiences and actions of human subjects (*structure/agency*), the interactional or micro level of society (*macro/micro*), and qualitative methods like interviews (*quantitative/qualitative*). This interpretation was sustained in the close reading and hand-coding of typical dissertations in this style (as presented in the previous section), where the style was characterized as fo-
cused on empirical problems and human agency. For the *person-oriented style*, *Lund* (38%), *Gothenburg* (31%), and *Uppsala* (25%) score the highest number of dissertations, and *Stockholm* (11%) and *Umeå* (10%) the lowest.

However, the concept of a methodological divide becomes slightly less tangible when considering the three remaining styles. Beginning with the second-best candidate for the qualitative community, the *concept-oriented style*, the seemingly tiny style was interpreted to capture words associated with the humanities (*science/humanities*) and classical sociology, as well as a focus on theoretical analyses (*empirical/theoretical*) of sociology and the human being in society, drawing heavily on social psychology (*structure/agency*). Here only three universities contribute with dissertations: *Lund* (9%), *Uppsala* (7%), and, to a lesser extent, *Stockholm* (3%). Taking only these three universities into consideration, there might be a case to be made that the spatial methodological divide can be found in the full-texts as well, although it is smaller than in the abstract corpus. Continuing with the largest style of the full-text corpus, the *relation-oriented style*, was interpreted to be constituted by words for addressing sociological theory and aspects of power and politics. The close reading suggested that this style appeared to be somewhat centered on scrutinizing social strictures rather than human action, and on discourse analysis as a method for studying texts. This combination makes the encoding of the style slightly less straightforward considering the methodological divide, yet it seems fair to say that the main methodology of the *relation-oriented style* leans a bit more toward the *qualitative community* rather than the *quantitative community*. If one accepts this interpretation, this leaning seems to be reflected in the style’s spatial distribution since all universities have a fairly large share of dissertations, with *Uppsala* (42%) and *Umeå* (41%) representing the largest share, closely followed by *Gothenburg* (36%), and *Lund* (28%), whilst *Stockholm* (24%) shows the smallest.

The last style, the *circumstance-oriented style*, was suggested to be the less distinct of the five styles since it did not seem to be directly linked to words associated with either qualitative or quantitative methodology and was understood to focus neither on theoretical or empirical matters, nor social structures or human agency. However, the results of the hand-coding based on close reading presented in the previous section, did unravel a pattern that the dissertations deemed to be typical for the *circumstance-oriented style*, due to their position in the stylometric network, were all applying the methodology of ethnography and studying meso-level phenomena. Thus, if one accepts that this style should therefore be seen as a closer associate to the *qualitative community* in comparison to the *quantitative community*, there is less pronounced but still noticeable distinction between the higher scores of *Uppsala* (20%) and *Lund* (18%), on the one hand, and *Stockholm* (11%) and *Umeå* (12%) on the other, with *Gothenburg* (15%) taking a middle position.
If reflecting on the distribution of all five styles vis-à-vis their content, it is clear that some form of spatial methodological division is to be found on the global level of the full-text corpus and, at the same time, the university corpora seem to express internal ruptures of their own. For the primarily quantitatively coded dissertations, it is the universities of Stockholm and Umeå that seem to have produced the greatest number of dissertations in the influential styles (i.e., the distribution-oriented style and the relation-oriented style). While both were interpreted to focus on social structures, the latter was suggested to put a larger emphasis on empirics, and the former on theory. Expanding on this pattern, a rift seems to emerge within the dissertations originating from two university corpora that allocates empirical dissertations on one side and theoretical ones on the other. Comparing the counts for the primarily empirical styles – the distribution-oriented style, the person-oriented style, and half of the circumstance-oriented style – with those of the most theoretical styles – the relation-oriented style, the concept-oriented style, and the other half of the circumstance-oriented style – Stockholm (67.5% vs 32.5%) appears more empirically inclined, with Umeå (53% vs 47%) more theoretically so.

For Lund and Uppsala, the two universities interpreted to produce dissertations that exemplify the qualitative community, the relation-oriented style with its focus on social structures also makes up the theoretical side of the divide and, on the empirical side, we find the person-oriented style which, in contrast, is interpreted to be focused on the experiences and actions of humans. Indeed, an interesting finding in itself is that we can note a strong presence of the person-oriented style, conceived to be constituted by words for qualitative methodology, and what seems to be an almost complete absence of full-texts associated with the distribution-oriented style, seen as a representative of quantitative methodology. The relation-oriented style of the full-text dissertations analyzed is further accompanied by relatively large shares of the small group of dissertations making up the concept-oriented style, characterized by an even more theoretical language and the absence of words found in empirical studies. Thus, one can sense a stylistic separation in dissertations defended in Lund and Uppsala between using a vocabulary addressing abstract societal entities and the experience of concrete persons.

In addition, we further find a sufficient population of dissertations sharing words comprised in the circumstance-oriented style, proposing a third, bridging type, where cultural phenomena – that can be understood as intangible ideas manifested in tangible practice – on the meso level are in focus. Again, we can find a division between the share of empirical and theoretical dissertations in each university corpora. When comparing the counts, Lund (54% vs 46%) comes out as the university that has produced the greatest number of dissertations with a more empirical focus and Uppsala (41% vs 59%) with a more theoretical one. The dissertations stemming from these
two universities are further differentiated in terms of subjectivist methodology – the person-oriented style, the concept-oriented style, and half of the circumstance-oriented style – and objectivist methodology – the distribution-oriented style, the relation-oriented style, and half of the circumstance-oriented style – where the former is favored by Lund (56% vs 44%) and the latter by Uppsala (42% vs 58%).

The last university, Gothenburg, takes on the same mediating role as in the previous chapters by displaying the most even stylistic distribution, with its greatest share of dissertations found within the person-oriented style and the relation-oriented style, as well as the circumstance-oriented style, like Lund and Uppsala, but also a sizable number of full-texts categorized as adhering to the distribution-oriented style, like Stockholm and Umeå. The only exception is the concept-oriented style, in which Gothenburg has no dissertations, and the university expresses a small leaning toward an empirical rather than a theoretical focus (56.5% vs 43.5%), and emphasis on social structures rather than human agents (61.5% vs 38.5%).

When combining the five universities on both sides of the so-called methodological divide proposed in the previous chapters, the analyses of full-texts sociology dissertations presented here (rather than abstracts as well as styles, rather than word constellations and thematic arrangements) seem to propose a more fine-grained spatial distribution. In simplified terms, the patterns exposed in this chapter suggest that the sociology dissertations hereby analyzed are not only dispersed over a methodological axis constituted by a qualitative community and a quantitative community but also an epistemological axis made up of an empirical community and a theoretical community. Looking at the dissertations produced in the quantitative-attuned universities (Stockholm and Umeå), we find a tendency to study social structures in society as either empirical populations or theoretical entities like discourses. For the dissertations produced in the qualitative-attuned universities (Uppsala and Lund), the opposition seems to be between depicting the social world through concrete experiences of people or abstract theories of humans.

Worth noting, though, is that while Stockholm and Umeå scored the highest number of sociology dissertations in the distribution-oriented style, they also had a good share in the person-oriented style, which is also interpreted to be an empirical style but based on qualitative instead of quantitative methods. Further, the two universities had an even larger share of dissertations in the relation-oriented style, a style deemed to be theoretical and focused on social structure but with a more qualitative than quantitative methodology. Correspondingly, Uppsala and Lund both had high scores in the relation-oriented style, which is not interpreted to be a clear-cut example of qualitative methodology, as well as a few percentages in the distribution-oriented
style, which is deemed to be the most clear-cut example of quantitative methodology. Thus, it seems reasonable to conclude that the spatial methodological divide expressed in the abstract corpus is less clear-cut in the full-text corpus. This is interesting, considering what the aim of this dissertation is since – although the analytical chapters addressed very specific research questions – the dissertation’s overall aim is the question of what computational analyses of dissertations (both abstracts and full-text versions of them) suggest about the sociological imagination and the paradigms that characterize the discipline that is sociology. Thus, although this chapter is not addressing this aim, there are results here that hint at where the final discussion of this dissertation will eventually lead.

To conclude this section on the temporal and spatial distribution of the five styles, the first finding to address is that the five styles mapped out in the full-text corpus throughout this chapter seem to be contingent on both time and place, yet the linguistic styles were more separate in the latter case. Indeed, when investigating the relationship between the linguistic styles the dissertations were assigned to, and the years they were defended, we found that the two empirical styles – the distribution-oriented style and the person-oriented style – occurred rather consistently throughout the time period. In comparison, the circumstance-oriented style showed a small ascending trend, and the concept-oriented style a clear descending trend, only to disappear completely a few years after the turn of the 2010s. While the relation-oriented style expressed a steady presence in general, it had a minor decrease in popularity in the 2010s relative to the 2000s. The two latter styles, the concept-oriented style and the relation-oriented style, are interpreted to pick up dissertations with a theoretical focus, and seen in this light there appears to be a regression in a theoretical style of writing sociological knowledge in favor of an empirical style. This trend was, to some extent, also traceable in the thematic analysis of the abstract corpus, where the more theoretical topics had their heydays in the 20th century and weakened in popularity throughout the 21st century. When considering that the person-oriented style and the concept-oriented style were construed to have a clearer focus on human agency than social structures, and that the opposite was interpreted to be the case for the distribution-oriented style and the relation-oriented style, there is also some evidence for a decline in the latter pole in favor of the former pole in the dichotomy of structure/agency. This pattern was not found in the structural topic model of the abstract corpus.

Continuing with place, the general interpretation put forward is that the university corpora in full-text are even more clearly divided in terms of style. While the results can be interpreted in light of the methodological divide that was mapped out in the abstract corpus, the overarching patterns appeared to be more complex than the ones unraveled in the previous chapters. However,
a more comprehensible and nuanced analysis beyond this methodological dichotomy of *quantitative/qualitative* was reached by, in addition, sorting the linguistic styles according to the epistemological dichotomy of *empirical/theoretical* and the ontological dichotomy of *structure/agency* discovered previously in this chapter. On the methodological and the ontological side, sociology dissertations defended in *Stockholm* and *Umeå* were found to be most strongly associated with the *quantitative* and the *structure* poles, but both seemed divided in terms of epistemology, where dissertations stemming from *Stockholm* appeared to be more strongly associated with empirical sociology and those from *Umeå* with theoretical sociology. For the two qualitatively coded universities, *Lund* and *Uppsala*, their university dissertation corpora seemed to imply a double rupture, where the former looked slightly more related to the *empirical* and the *agency* poles and the latter to the *theoretical* and *structure* poles. Stylistically, dissertations defended in *Gothenburg* embody a bridging position, as was also the case in the previous analyses of the abstract corpus. This is supported, among other things, by showing roughly three times the score in the *distribution-oriented style* compared to *Lund* and *Uppsala*, as well as roughly three times the score in the *person-oriented style* in comparison to *Stockholm* and *Umeå*. While situated in the middle of the methodological divide, *Gothenburg* appeared to be slightly more inclined to the *qualitative, empirical*, and *structure* poles than the *quantitative, theoretical*, and *agency* poles.

**Closing Thoughts on Linguistic Styles**

To investigate the three research questions posed in this dissertation from the angle of stylometry, this chapter launched an investigation of the full-text corpus – comprising 380 sociology dissertations defended between the years 2000 and 2019 – to identify stylistic patterns in how sociological knowledge is presented in the written form that is dissertations. Beginning with the first research question, exploring the so-called fragmentation of sociology, it was suggested that sociological knowledge is patterned according to linguistic styles that recur over time. These linguistic styles were modeled from the actual texts constituting the corpus based on computations of linguistic presence. The process began by computing bootstrapped consensus trees on the full-text corpus whereby a stylometric network was generated. By adding a modularity algorithm, five communities of various sizes were found that, with the help of contrasting Burrow’s z-scores, were interpreted to relate to five separated linguistics styles associated with certain types of word usage.

An attempt to capture the main tendencies of the linguistic styles was made by assigning each style a label based on its z-scores: the *relation-oriented*
style (n=130), the person-oriented style (n=84), the distribution-oriented style (n=98), the circumstance-oriented style (n=54), and the concept-oriented style (n=14). The key aspects of the differences between the linguistic styles were further explored through close reading and hand-coding of the five most typical full-texts for each style, which were selected by singling out the dissertations with the highest degree within its assigned community and the lowest degree outside that community in the stylometric network. It was suggested that the linguistic styles of these dissertations not only diverged in their empirical themes and methodologies but also seemed to differ in their basic assumptions concerning epistemology and ontology. The proposed existence of linguistic styles seems to suggest that sociological knowledge, as it is presented in the full-text corpora, is not overly fragment-ed. Rather, there seem to be stylistic patterns that shape and ‘hold’ the production of sociological knowledge together. Given the fact that the stylometric analyses performed are based on the most frequent words in the corpus, these linguistic styles appear to articulate uniting aspects of how sociological knowledge is written, which exist ‘underneath’ the level of thematic specialization (as expressed in the two previous chapters exploring word constellations and topics in the abstract corpus).

Continuing with the second research question, which explores paradigms, the prevailing dichotomies of sociology turned out to be effective tools for analytically separating central aspects of linguistic style. The analyses presented in the chapter included almost all nine dichotomies, yet three dichotomies turned out to be the most important for separating the linguistic styles: the methodological dichotomy of quantitative/qualitative, the epistemological dichotomy of empirical/theoretical, and, lastly, the ontological dichotomy of structure/agency. While the person-oriented style and the distribution-oriented style were both understood to be seated in the empirical pole, the former seemed to resonate with the poles of qualitative and agency and the latter with those of quantitative and structure. In contrast, the concept-oriented style and the relation-oriented style were both construed as examples of linguistic styles with a theoretical emphasis and more qualitative-attuned than quantitative-attuned. Yet, it was further suggested that dissertations in the two styles diverged in that a concept-oriented style tended to focus on human agency and a relation-oriented style on social structure. Lastly, dissertations with a circumstance-oriented style were found to be associated with both poles of many dichotomies. However, based on the close reading with hand-coding of the five most typical dissertations in the style, it was suggested that all were ethnographies of phenomena on the me-

126 In network analysis, ‘degree’ refers to the measure of the total number of links or edges connected to a particular node.
so level, which is at least some evidence for a leaning toward the empirical pole and the qualitative pole.

In other words, the five linguistic styles appear to reiterate a combination of dichotomies that cannot easily be sorted into a single master dichotomy. Indeed, while full-texts with a person-oriented style and a distribution-oriented style were presented to be representative of the qualitative community and the quantitative community, respectively, corresponding to the so-called methodological divide mapped out through the analyses performed on the abstract corpus as presented in Chapter 6 and Chapter 7, the three remaining styles – the concept-oriented style, the relation-oriented style, and the circumstance-oriented style – seem to suggest that the full-texts follow a much more multifaceted configuration. Thus, these results appear to differ from the ones presented in the two previous analytical chapters that were based on the abstract corpus, and differ also from results presented in previous studies as alluded to in the literature review in Chapter 2.

Following the paradigm theory applied in this dissertation, the act of repeating the same dichotomies within a linguistic style can be interpreted as forming crude replicas of a paradigm. Thus, the way sociological knowledge is articulated in the full-text corpus is not interpreted to reflect a dual-paradigm science. Rather, it seems the five linguistic styles mapped out in the full-text corpus reflect five different paradigms. In this sense, if looking exclusively at what the results of this chapter suggest, it might be plausible to describe the full-text corpus as conveying a picture of sociology as a multiple-paradigm science. This tentative interpretation of the full-text corpus will be investigated further in the next chapter by utilizing word embeddings, which is a computational technique for representing words in a vector space where proximity entails semantic similarity.

With the five linguistic styles at hand, it was also possible to continue the investigation of the third research question, which focuses on the social conditioning of sociological knowledge. By sorting the full-texts into corpora of the year when, and the universities where, the dissertations were defended, the analyses performed on descriptive statistics showed that the linguistic styles were unequally distributed over time and place. Dissertations with a distribution-oriented style, person-oriented style, and relation-oriented style seemed to occur rather consistently over time. In contrast, the concept-oriented style expressed a clear descending trend and the circumstance-oriented style expressed a small ascending trend among the dissertations. Regarding spatiality, the relative shares of the distribution-oriented style were dominated by the dissertations defended in Stockholm (51%) and Umeå (37%), while the opposite was the case for the person-oriented style, Lund (38%), Uppsala (25%), and Gothenburg (31%). Further, Uppsala and Lund
had the highest relative shares for the few dissertations that articulate a concept-oriented style (7-9%). Lastly, the relation-oriented style and the circumstance-oriented style of sociology dissertations hereby analyzed were more fairly distributed among the universities, with Uppsala (42%) and Umeå (41%) expressing the highest scores.

When viewing the temporal dimension of the linguistic styles in the light of the sociological dichotomies, the results of the analyses presented in this chapter pointed towards the 2000s as a period more associated with the theoretical pole and the agency pole, while the 2010s seemed more related to the empirical pole and the structure pole. In a corresponding fashion, the spatial investigation suggested that while dissertations defended in Stockholm and Umeå were mostly constituted by full-texts categorized in styles related to the quantitative pole and the structure pole, the dissertations from the former university appeared to be characterized more by an empirical focus and the latter with a theoretical focus. In comparison, the two universities understood to be situated at the qualitative pole – Uppsala and Lund – differed for both the other poles, where Uppsala was deemed to be more attuned to the theoretical and social structure and Lund to the empirical and human agency. Lastly, dissertations defended in Gothenburg were slightly more associated with the qualitative pole, yet scored significantly higher in the distribution-oriented style in comparison to the two qualitative-attuned universities, and were therefore suggested to be a little closer to the theoretical pole and the agency pole, respectively.

Given the cumulative design of this study, it is worth discussing the congruence between what the analyses of abstracts suggested and what the full-text inquiry presented in this chapter has shown. It was proposed that the patterns found in this chapter implied a break with the narrative of the methodological divide as the main locus for unraveling sociology dissertations. This suggests that there are potential consequences of analyzing corpora on the basis of abstracts or full-texts. This result poses a fundamental question to the computational studies of sociology examined in the literature review of this dissertation, namely whether abstracts can be deemed proper empirical material for detecting patterns in sociological knowledge. As shown in the literature review chapter (Chapter 2), several studies conducted by leading sociologists within computational social sciences have based their data solely on abstracts, titles, and keywords (e.g., Moody and Light, 2006; Seale, 2008; Zougiris, 2019; Moody et al., 2022).

Thus, the underlying question hereby alluded to – which is also part of the reason why this dissertation’s aim reads as it does – is to what degree an abstract serves the function of summarizing a paper or book, and to what extent it should be deemed to be a written presentation of how a dissertation
(as an example of a product of sociological knowledge) ought to be read, which can be in great conflict with what the full-text is actually addressing. Nevertheless, one more analytical chapter remains. Because of the cumulative study design utilized here, it is in the concluding discussion of Chapter 10 that the full answer to each of the three research questions will be given. The next chapter will study the distributional meanings of sociological knowledge in the full-text corpus rather than the styles of writing sociological knowledge, which has been the focus of this chapter.
9 Distributional Meanings

This chapter expands on the line of reasoning offered in the three previous analytical chapters (Chapter 6, Chapter 7, and Chapter 8) by approaching the problem of sociology from a new angle. This move entails taking a step beyond the prevailing understandings of sociology addressed in the three research questions guiding this dissertation, which were generated after reviewing the theoretical models employed in previous empirical studies within the sociology of sociology (see Chapter 2). The exploration presented in this chapter is based on the notion that the foundation of the crisis of sociology might not necessarily be found in the fragmentation of sociological knowledge or the absence of a paradigm, but sought instead in the ‘meanings’ that are attached to the foundational concepts of the discipline. As presented in the theory chapter (Chapter 3), Ritzer’s (1975, p. 157) work on paradigms in sociology argued that a paradigm rests on “a fundamental image of the subject matter”. In other words, a paradigm in sociology encapsulates some form of shared imagination of what sociology is.

The approach applied in this chapter is based on finding whether divergent meanings of the fundamental image of sociology’s subject matter are articulated in the full-text corpus. This corpus comprises 380 full-text sociology dissertations defended and published between the years 2000 and 2019. Yet, compared to the previous chapter, the locus is moved from linguistic styles to mapping ‘meanings’ of sociological knowledge. To make this shift, one has to move away from treating a document as a long row of disentangled words, the so-called bag-of-words model, to take into account the neighboring words of every single word occurring in each book for the whole corpus, which gives us vectors of features often referred to as word embeddings. To investigate the temporal and spatial conditioning of what type of ‘fundamental image of sociology’ the dissertations articulate, the variables for the time and place of the doctoral defense will once again be utilized. In addition, to explore the relationship between the dissertations’ stylistic features and the meanings being attached to fundamental concepts in sociology, the five linguistic styles generated in Chapter 8 will also be utilized as metadata.
Word embeddings rely on a fundamental assumption regarding lexical semantics referred to as the distributional hypothesis,\textsuperscript{127} often popularly summarized in the phrase “You shall know a word by the company it keeps!” (Firth, 1957, p. 11), which builds on the later Wittgenstein’s (2009, p. 25e) notion that “the meaning of a word is in its use in the language”,\textsuperscript{128} An idea that is commonplace among contemporary computer scientists – which was also proposed by pioneering computational linguist Masterman (1970) decades ago – is that representing a word’s context is representing its meaning. The logic goes something like this. Words are first converted into vectors based on their probability to occur next to other words. This gives us a vector space where words sharing the same specific word neighbors are given adjacent positions. Then, semantic similarity can be estimated based on how close two words are to each other in this vector space, which, in simplified terms, is a representations of how likely it is that these words would occur within the same context of words. In simplified terms, this procedure would give us the meaning of the words. However, to argue that machines alone can extract meanings from texts – something that is often implicit and seldom differs regarding the cultural script through which it is deciphered – is a philosophical position that probably appears irrational to most sociologists. Thus, in this chapter, the fundamental meaning of a word (if there is such a thing) is never discussed. Nevertheless, since the task set out here is finding out if there is any potential leverage on computational methods for studying texts, the phrase distributional meaning is employed to delineate the analyst’s interpretation of a word embeddings connotation based on its context, that is its position in vector space and the word embeddings that are its closest neighbors; see Appendix E for the technical details of how the word embeddings utilized in this chapter was generated.

The presentation of this analysis follows four steps. First, the word embedding model utilized in the analyses is presented along the general idea behind how the word embeddings are interpreted. Thereafter, the distributional meanings of the word embeddings sociology and society, respectively, are investigated as well as how they are distributed over time, space, and style. Lastly, we return to the divisions of sociological knowledge explained in the theory chapter to see how they are expressed in vector space and if they follow any specific temporal, spatial, or stylistic patterns.

\textsuperscript{127} For a more in-depth explanation of the distributional hypothesis, please see the methodology chapter and Appendix E.

\textsuperscript{128} “Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache.” (Wittgenstein, 2009, p. 25)
Training and Analyzing Word Embeddings

This short section follows a simple technical description of how the word embedding model was trained, as well as how it is utilized in the analysis. The more technical and mathematical descriptions of the methods employed in this chapter are offered in Appendix E. The computational technique applied in this chapter for generating word embeddings is called skip-gram word2vec, which is a prediction-based unsupervised machine algorithm that can “predict words within a certain range before and after the current word” (Mikolov et al., 2013, p. 4). For the purpose of the analyses performed in this chapter, a skip-gram model gives the opportunity to predict the probability of all possible contexts (or surrounding words) of a word of interest and, thus, to map out their distributional meaning. In practical terms, the word embeddings have been generated by first taking the 66,000 unique words occurring in the sociology dissertation full-text corpus and locating each word’s closest neighbors in every sentence where it is used (a sentence is here defined as a word window of eight on each side of the target word). These values are then converted into sequences of numbers or vectors of features called word embeddings. Each of the 66,000 word embeddings have 300 dimensions that hold information on their distance to the other words, giving us a vector space constituted by almost 20,000,000 elements.

Again, following the distributional hypothesis, closeness or remoteness in the modeled vector space is seen as equivalent to closeness or remoteness in distributional meaning. Besides finding out the distributional meanings of words, the model can also be utilized to calculate how close the different documents, in this case, sociology dissertations, are to specific word embeddings in vector space by leveraging a method developed by sociologists called Concept Mover’s Distance or CMD (Stoltz and Taylor, 2019). A much-simplified and technical definition of this method is “the minimum cost that a document’s embedded words need to travel to arrive at the position of all the words in an ideal ‘pseudo document’ consisting of only words denoting a specified concept” (Stoltz and Taylor, 2019, p. 294). CMD is repeatedly utilized in the analyses of the chapter to find the extent to which the sociology dissertations analyzed here engage with different concepts, and

129 Other famous frameworks for generating word embeddings were run and the results were compared before deciding to go with Word2vec’s skip-gram, since it performed best on the tasks executed on this relatively small dataset, namely Word2vec CBOW (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and BERT (Devlin et al., 2018).
130 Only words occurring three times in the whole corpus are included to remove unnecessary noise such as typos from the model. The original count of unique words was a little over 174,000.
131 The corpus is made up almost 23 million words distributed over 380 sociology dissertations in full-text.
how this engagement differs in relation to temporality, spatiality, and style. In both cases – the word embeddings’ closeness to each other and to documents, respectively – the distance metric is based on cosine similarity. Thus, with the help of the word embedding model at our disposal, we can head on to locate the distributional meaning of sociological knowledge as it is actually used in the corpus.

Distributional Meanings of Sociology

The overarching backdrop to this dissertation is the discipline’s struggle to define what sociology is. Thus, to begin our exploration, we will model the distributional meaning of the word sociology based on how it is used in the full-text corpus. This is done by scrutinizing the neighbors of the word embedding sociology, which make up its context. This section is fully devoted to analyzing sociology as well as investigating what sociology dissertations contribute to the distributional meanings of this particular word embedding. The latter is conducted by comparing the distance between sociology and the familiar categories time, place, and linguistic style in the vector space generated by the model described in the previous section. As usual, time is represented by the year a dissertation was published and place by the university that published it. Linguistic style is based on the stylometric categories that were generated on the basis of the dissertations’ textual content in the previous chapter. To map the context of sociology, we will look into its five closest neighbors and, to understand their distributional meanings, also look into their five closest neighbors, as well as the five closest of their neighbors. By adding up all these word embeddings and converting distances to edges (weights), we can generate a large word network that can be interpreted to find distributional meanings of sociology based on its context. Such a network is presented in Figure 23, where the word embeddings are represented by orange nodes and the distance between neighboring words in vector space are edges.
Figure 23. Network based on the text2vec skip-gram algorithm, here modeled on the full-text sociology dissertation corpus where the word embedding “sociology” is targeted, as well as its five closest word neighbors and their five closest neighbors. Edge thickness indicates closeness in vector space and node size represents degree.

Taking a bird’s-eye view on Figure 23, we can see that it is naturally centered around the word embedding sociology, our empirical object of study, and what appears to be four main sets of words. Starting with the top part of the network, the top-left side is constituted by terms for common roles within academia in general, like researchers and scholars, and for sociology in particular, like sociologists, and the top-right side encapsulates words related to the sciences and the humanities, primarily disciplines within the social sciences. On the bottom part of the network, we find, to the left, words related to theories and theorists and, to the right, terms drawn from philosophy, primarily the branches of epistemology and metaphysics. Thus, the overarch-
ing context of sociology, as drawn from a human interpretation of the computed word embeddings, seems to comprise its place in academia – both where the discipline of sociology is positioned and what the role of the sociologist is related to – and its conceptual content – both in terms of generating and applying theories, as well as discussions on its philosophical underpinnings.

To justify the distributional hypothesis underpinning this dissertation, we cannot stay at this interpretation of sociology. Rather, we would have to unpack the context for each of the closest neighbors to the word embedding sociology to get an idea of their meanings within the corpus and apply these to a more nuanced understanding of the usage and meaning of sociology. In Figure 23, the closest neighbor to sociology is sociological, followed by science, philosophy, sociologists and sciences. In other words, when addressing sociology in the corpus, it is most often in the same context as one or several of these five words. Therefore, we need to get an idea of the connotations of these word embeddings to understand the meaning of sociology as it is used in the corpus.

Beginning with sociological, the closest neighbor to sociology in vector space, its five closest neighbors (besides sociology) are theoretical, theory, sociologists, theories, and classical. Thus, sociological is found in the same context as other generic words for sociology – sociology and sociologists – as well as a collection of words for theory – theoretical, theory, and theories – and, as a particular token, classical. Beginning with how to understand one of the principal parts of the context, the theory words, we find, unsurprisingly, three words with the same stem that are interconnected with each other – theoretical, theory, and theories. Opening with the noun in its singular form, theory, and plural form, theories, these are related to the word embeddings concept and concepts, which are assumed to be the building blocks of theories (empirical/theoretical) and, thus, share similar connotations. Further, the word embeddings theory and theories are also connected to what appear to be generic words for taking a point of view – approach, approaches, and perspectives – as well as critique, which entails a more critical position. The tendency that theory appears to be used synonymously with a point of view is deemed interesting in relation to other potential ways of conceptualizing theory. For instance, the natural scientific conceptualization of theory is often described as a synthesis of facts and hypotheses that explain natural phenomena rather than a perspective as we would find in the humanities (cf. Parsons, 1970). Thus, the context of sociological is here understood to be more related to the latter disciplines than the former (science/humanities).
The word embedding theoretical carries a wider context than the other theory words in the sense that it is related to words ranging from conceptual through analytical and methodological to empirical, entailing that theoretical is not solely deployed in relation to abstract ideas but also to concrete research practice of empirical inquiry (empirical/theoretical). Another last word embedding in the immediate context of sociological is classical, which, besides being connected to sociology words, is distributionally adjacent to words with what appear to be similar connotations. These include a word with the same stem, classics, the antonym modern, and the specific term marxist, presumably referring to followers of the intellectual program and theory laid out by one the main classics in the discipline of sociology, Karl Marx. Thus, classical appears to be related to classical works of sociology and to some extent related to theory. Taking into consideration all theory words related to sociological, the distributional meaning of sociological theory looks to be primarily about concepts, as well as taking a conceptual point of view, and involves discussions of classics. Further, theory to some extent appears in the same context as analytical, methodological, and empirical concerns.

A close neighboring word to sociological is sociologists, which is also, as mentioned above, one of the most adjacent words to sociology. This token, sociologists, is primarily connected to the already mentioned generic words for sociology, including its singular form, sociologist. The latter is adjacent to both philosopher and researcher, entailing a tension between what other roles the role of the sociologist is discussed in relation to. This pattern is enhanced when zooming in on the other four words adjacent to sociologists, which share similar connotations: researchers, scholars, theorists, and scientists. These words seem to signify four roles that encapsulate different understandings of “what sociologists do”. On the one hand, we find scholars and theorists – the first reaching out to scholarship and commentator as well as researchers, and the second to thinkers as well as the theory words discussed above. The two first roles, thus, seem to be about engaging in intelligent and scholarly endeavors (science/humanities), where scholars is the word embedding drawing more towards research. On the other hand, we have the word embedding researchers that is only connected to other words with the same stem – research and researcher – and scientists, which also have close neighbors with the same stem – sciences and science – as well as the more particular term economists, entailing that this discipline is addressed in the context of science. Thus, the last two roles appear to be more about conducting research and engaging in scientific activities (science/humanities). We can perhaps assume that researchers and scholars are less laden roles than theorists and scientists since the two former form a bridge between the two latter. Nevertheless, the answer to the question of “what soci-
ologists do” seems to encapsulate a range from developing theory and traditional scholarly pursuits, presumably reading and writing literature, to researching and scientific endeavors like that of economists, known for their appreciation for quantitative data and statistical methods. Thus, while the word embedding sociological is primarily associated with theory (empirical/theoretical), sociologists comes with a more wide-ranging context occupied by as many humanistic-scholarly (science/humanities) undertones as scientific-research ones (science/humanities).

Besides sociological and sociologists, three remaining words make up the five closest neighbors to sociology, namely the interconnected word embeddings of science, sciences, and philosophy. The two first word embeddings are linked to two other word embeddings sharing the same stem: scientists and scientific. While the former, scientists, naturally take us back to the role of researchers but also of economists, which, as was discussed above, we can assume is talked about in the same context as scientists, the latter, scientific, is situated near knowledge and epistemological, interpreted as referring to what counts as knowledge and methods for gaining knowledge, as well as philosophical. Further, science and sciences are connected to humanities, which in turn is linked to ethnology and faculty, as well as disciplines, that is connected to discipline, subfields, and economics, as well as psychology, sharing its context with other words with the stem “psycho”, namely psychological, psychologists, and psychodynamic. Thus, the distributional meaning of the science words (science/humanities) in the corpus appears to be associated with philosophy, primarily epistemology, words for generic academic entities, and the three social sciences – sociology, economics, and psychology – as well as the ethnology branch of the humanities (science/humanities).

Continuing with the last immediate neighbor of sociology as well as of science and sciences, that is philosophy, its closest neighbors are two interconnected word embeddings with the same stem, philosophical and philosophers. The first word embedding is adjacent to other names for roles centered on intellectual work – philosopher, thinkers, and logician – and the latter, which has already been discussed in the context of the science words, is linked to scientific and epistemological but also the word metaphilosophy, alluding to a philosophy of philosophy. Further, philosophy is in close adjacency to two other interconnected tokens: meta-

---

132 If we think back to the introduction chapter it is perhaps curious to note that teaching does not appear in the words that this data corpora suggests as associated with the word sociologist. This raises the question of whether “teacher” would arise as a word closely related to sociologist if we were to have analyzed a corpus of articles published in the journal called Teaching Sociology. Future research could perhaps investigate this, and of course whether these word embeddings are nation-specific.
physics and positivism. Adjacent to metaphysics, we find the related word embeddings metaphysical and hypostasis, which is here understood as the word traditionally used for addressing the metaphysical substance upholding reality, as well as deconstruction that is interpreted to mean the theory and method that accentuate the instability of reality and meaning (entailing a negation of essences). Thus, the word metaphysics is here conceived to entail the branch of philosophy dealing with finding the fundamental nature of reality (science/humanities). In turn, positivism is linked with three word embeddings sharing the same stem — positivisms, positivist, positivists — as well as empiricism and logical. The former word, empiricism, is comprehended to entail the philosophical position holding that knowledge ought to be based on experience derived from the senses. The latter word, logical, is rather generic but in this context of words it is understood to reference logical positivism or logical empiricism, a philosophic movement upholding the position of science and analytical philosophy (science/humanities). Looking at the distributional context of word embedding positivism, the word is interpreted to mean the philosophical system centered on the belief that knowledge can only be generated via scientific method, logic, and mathematics. Since positivism usually entails a firm rejection of traditional metaphysics, we can speculate if the two words might be used to accentuate two different routes that sociology can take.

An immediate interpretation of the philosophy words and the science words could be that they represent a tension between dissertations dedicated to either a more humanistic or a more scientific sociology (science/humanities). However, it is worth noting that, except for philosophers and scientists, all words with the stem “scien” are adjacent to their corresponding word with the stem “philosoph” and vice versa. In other words, terms for science and philosophy, respectively, are not isolated from each other in the corpus but rather have a high probability to co-occur in the same sentences. In other words, it does not appear to be the case that some dissertations are solely concerned with science while others only deal with philosophy.

With a detailed view of all words constituting Figure 23, we can return to the initial overview of the network that emphasized sociology’s place in academia and conceptual content. It is clear that sociology is at times discussed concerning some other social science disciplines like psychology and economics, but primarily the usage of the word sociology seems to be synonymous with addressing theoretical concerns (empirical/theoretical). Following the main tenets found in many theories in the sociology of sociology, a tempting interpretation of the remaining distributional meanings is that they are divided into two camps with different understandings of what sociology is and where it should go. On the one hand, we have proponents for a type of sociology drawing on a scholarly tradition associated with the humanities
(science/humanities) and primarily philosophy that addresses big questions like the metaphysics of society, and on the other, those closer to the idea of positivism and seeing sociology as a science conducting empirical research (science/humanities) and, thus, closer to the likes of economics and psychology may stress another sociological type. In simplified terms, we can imagine this opposition to be the same two opposing cultures, described by Wolf Lepenies (1988), that have been struggling for centuries over the questions of the human condition in modern society and how it ought to be studied, with literature or humanities on the one side and science on the other. Yet, as was briefly addressed in the paragraph above, the word embeddings alluding to science and philosophy, respectively, are not disconnected – which would entail that they do not appear together in the text – but embedded in the same context. Thus, it appears to be more probable that science and philosophy would tend to occur in the same sentences addressing both entities in tandem, rather than different sentences engaged in separate discussions.

The relationship between science and philosophy in vector space can be explored further since word embeddings are made up of vectors, so you can simply remove the context most associated with one word embedding from another. If, for instance, we remove word embeddings that traditionally have had methodological connotations within sociology from the word embedding sociology, the closest word embedding becomes philosophy (see footnote for the methodology word embeddings considered). This is rather intuitive since sociology without any empirical inquiry tends to draw heavily on philosophy (a theoretical mode of sociology represented by the concept-oriented style found in the previous chapter). When removing the context of philosophy from that of science, the nearest word embedding becomes engineering (science - philosophy = engineering), followed closely by technology. This seems to make sense since one can imagine that if you remove the most philosophical (i.e., the abstract or conceptual) words about science you end up with engineering and technology, which are often conceptualized as concrete, practical, or applied versions of science. The opposite calculation gives you hegel (philosophy - science = hegel) and thereafter derrida; Hegel and Derrida are two thinkers who can be understood as representatives of continental philosophy, a term invented by the analytical movement in philosophy to rule out all philosophies not adhering to the form of reason and analysis associated with analytical philosophy and logical positivism (cf. Critchley, 2001). This interpretation appears to be sustained when adding the word embeddings together and combining their

---

133 The following calculations were made by removing a word embedding from sociology and showing its nearest neighbor: sociology - method = philosophy; sociology - data = philosophy; Sociology - quantitative = philosophy; sociology - qualitative = philosophy.
contexts since the model then gives us perhaps the most outspoken program in favor of science within the philosophical discipline: positivism (philosophy + science = positivism; science + philosophy = positivism). Thus, even though science and philosophy are very close neighbors to both sociology and each other in vector space, there are at least some dimensions (of the 300 constituting the word embedding model) that are signaling their differences.

To investigate the extent to which the analyzed dissertations are addressing sociology, theory, philosophy, and science, i.e., what seem to be the key themes constituting the distributional meaning of sociology, we can look at how close the dissertations are to each of the four. This is done with the method Concept Mover’s Distance described in Appendix E, which can compute how close the dissertations are to a set of word embeddings. The word embeddings together form what is called a “centroid” based on the average of their positions in the multidimensional vector space. For conducting this task, we will have to construct a tiny dictionary and compute a centroid for each set of word embeddings constituting the distributional meaning of sociology explained in the previous section. This entails one dictionary for sociology words, another for science words, a third for philosophy words, and, lastly, a fourth for theory words. To get a comprehensible understanding of how the dissertations are distributed to the centroids, we can apply the three metadata used in the previous chapter – the years the dissertations were published, the universities where they were published, and the linguistic styles – to investigate if the distributional meaning of sociology is associated with any temporal, spatial or stylistic configurations. By this technique, we can also continue our investigation of whether there is evidence for either a set of dissertations dedicated to philosophy and another to science, or if there is a single set of dissertations discussing both philosophy and science.

To begin, we can look at how time might impact to what extent the dissertations address the word embeddings found to represent sociology in the vector space generated by the word embeddings model. The temporal influence is here measured as the relative closeness of all dissertations published in

---

134 The dictionary forming the sociology centroid consists of the following word embeddings: sociology, sociologys, sociologies, sociological, sociologist, sociologists.
135 The dictionary forming the science centroid consists of the following word embeddings: science, sciences, scientific, scientist, scientists.
136 The dictionary forming the philosophy centroid consists of the following word embeddings: philosophy, philosophys, philosophies, philosophical, philosopher, philosophers.
137 The dictionary forming the theory centroid consists of the following word embeddings: theory, theorys, theories, theoretical, theorist, theorists.
different years between 2000 and 2019 to the semantic position of the sociology centroid, science centroid, philosophy centroid, and theory centroid, respectively. A visualization of this computation can be seen in Figure 24, where the x-axis depicts the year of publication and the y-axis the dissertations’ relative closeness (+) or distance (-) to a centroid in the vector space. There are four lines in the figure representing the annual conditional mean of the philosophy centroid (gold), the science centroid (blue), the sociology centroid (red), and the theory centroid (black). Each point indicates the relative closeness to a dissertation published in that year to a particular centroid. The points are colored in alignment with the centroids.\textsuperscript{138}

![Figure 24](image)

**Figure 24.** A time series of the distance between the dissertation full-texts and four dictionary-based centroids (philosophy, science, sociology, and theory) which are conceptualized to represent the distributional meaning of “sociology” in vector space.

In Figure 24, we can see that the dissertations, represented by points, take on high and low values throughout the period; however, the majority surround the zero value of the y-axis, indicating that they are neither close nor far

\textsuperscript{138} The points are jittered (i.e., randomly separated a tad) to avoid too much overlap.
away from the centroids in vector space. Therefore, as expected, the general pattern for the four lines is a relatively steady presence throughout the whole time period. In simplified terms, time appears to have no substantial influence on the distributional meaning of sociology in the corpus. However, there are a few aspects of the different lines worth exploring a little further. The most evenly allocated centroid in terms of time seems to be represented by the theory centroid, with its only exception being that it has the lowest distributional mean of all centroids during 2000-2002. A close candidate in terms of stability over time is the sociology centroid, which only shows a small drop and rise at the beginning, and the opposite pattern at the end of the time period. Thus, these two centroids appear to be slightly more steadily distributed over the years than the other two. In comparison, the science centroid is fluctuating more clearly throughout the time period, repeatedly going over and under the center of the y-axis. One noticeable observation with the science centroid is that it has one of its highest peaks around the year 2018 when the other centroids express a downward trend. Lastly, the centroid with the most frequent and strongest shifts appears to be found in the philosophy centroid, with its conditional mean value showing the highest and lowest value of all four centroids, the former during 2002-2007, with its peak around 2005, and the latter 2015-2019, with its all-time low in 2017. Thus, perhaps a valid proposition, which should be taken with a pinch of salt due to the overall low values, is that the philosophy centroid is associated with the earlier rather than the later parts of the time period.

While the four centroids are not identically distributed over time, they do seem to follow almost the same fluctuation pattern, which is expected since they are close neighbors in the static vector space modeled on all dissertations. The pattern appears to be most clear for the philosophy centroid and the science centroid, while the sociology centroid and the theory centroid look like they more often break with the pattern of the other centroids; this is particularly true for the latter centroid. One potential interpretation of why the philosophy centroid and the science centroid both share a similar temporal pattern, which was presented in the previous section and is here put forward with some hesitation, is that usage of science words and philosophy words in the corpus are interlinked but differ in the sense that science gains popularity vis-à-vis philosophy in the late 2010s. An equivalent interpretation with the same amount of reservation can therefore be put forward concerning the shared pattern of the sociology centroid and the theory centroid, where the latter is slightly more prevalent at the end of the time period.

With a rudimentary understanding of the relatively weak temporal influence on the distributional meaning of the word embedding sociology, we can move on to the aspects of style and place. To recall, place is measured as the university where the dissertation was published, and style by the linguistic
style category the dissertation was assigned to by the algorithm forming the stylometric bootstrapped consensus network. When the two metadata were compared in the previous chapter, it was suggested that while place seems to have been an important distributional factor in all analyses, particularly those performed on the abstract corpus, linguistic style (that is modeled on the actual content of the full-text corpus) might be a better alternative for mapping patterns of sociological knowledge on text data. To be able to continue exploring this methodological question, the two will be investigated in tandem when computing the influence of style and place on the four centroids.

Figure 25 is a visualization of the influence that style and place have on the distributional meaning of sociology, which is understood to be centered around the four centroids presented above: philosophy centroid, science centroid, sociology centroid, and theory centroid. The y-axis for each graph is a computation of how near or close a dissertation (represented by a symbol) is to the dictionary constituting a centroid in vector space, and the boxplots show the distribution of all members of either the spatial university category or the style category. If comparing all the centroids to each other, it seems like the sociology centroid and the theory centroid, on the one hand, and the science centroid and the philosophy centroid, on the other, share a similar distribution in both the university category and the style category. When looking at the distributions of the full-text corpus, sorted into five universities and the five styles, respectively, one overarching pattern emerging is that the university groupings have more dispersed distributions – i.e., have a wider span between the upper and lower quartile, and all universities have some dissertations reaching some of the highest and lowest values in the corpus for at least one of the centroids – but their mean values are to a greater extent closer concentrated than the style groupings.
Figure 25. A comparison of the dissertation full-texts, divided into groups of style and place, and the four dictionary-based centroids (philosophy, science, sociology, and theory) conceptualized to represent the distributional meaning of “sociology” in vector space.

For all four centroids, full-texts stemming from two of the universities, Lund and Umeå, have a mean value centered around zero with a positive value for one centroid, philosophy in the case of Lund, and theory for Umeå. Yet these two universities show a rather wide span between their highest-scoring dissertations and their lowest-scoring ones. In comparison, full-texts originating from Uppsala express little positive mean values, and full-texts from Stock-
holm and Gothenburg show slight negative mean values for all centroids as well as a little more concentrated distribution of dissertations along the y-axis. Although we might sense a separation of university groupings of full-texts between relatively positive Uppsala and relatively negative Stockholm and Gothenburg, it is hard to find a general division between all universities since, for instance, full-texts defended at Lund and Umeå score quite similarly. Concerning the previous chapters that showed a comparatively clear division between dissertations from Uppsala and Lund on one side and those deriving from Stockholm and Umeå on the other, this result is somewhat unexpected and suggests that not one university grouping of full-texts stands out drastically. In other words, like time, place does not appear to be a major factor for who is forming the distributional meaning of sociology.

As mentioned, the linguistic style category appears, in general, less scattered than the university category – with the most evenly distributed range found in full-texts categorized in either the circumstance-oriented style or the distribution-oriented style, while dissertations falling under the relation-oriented style and the concept-oriented style are the least evenly distributed. In between these four styles, we find the person-oriented style that has the majority of its dissertations concentrated within a rather small range but also includes a few outliers for each centroid. Comparing the linguistic styles to each other, they appear to form a descending pattern from left to right that starts with the relatively high scores of full-texts in the concept-oriented style, followed by the relation-oriented style and the person-oriented style, and ends with the circumstance-oriented style and the distribution-oriented style. It seems that it is the concept-oriented style and, to a much lesser extent, the relation-oriented style that house full-texts that are closest to word embeddings related to sociology (theory, science, and philosophy). In contrast, full-texts of the distribution-oriented style scored low for all centroids and, as expected, particularly the philosophy centroid, but slightly surprisingly, it also has relatively low scores in the science centroid, which is interesting since the distribution-oriented style is construed to constitute full-texts with a high share of scientistic terms. The tendency that dissertations categorized within the tiny concept-oriented style, as well as some dissertations falling within of the relation-oriented style, are such high scorers in all four centroids alludes to the interpretation that there might be a set of full-texts representing the main distributional meaning of the word embedding sociology.

Based on the presented interpretation of Figure 25, where the mean values of the linguistic styles differ more clearly from each other in relation to the universities, it seems plausible to suggest that style might be a better indicator than place for finding who contributes to the distributional meaning of sociology. Further, the linguistic styles scoring high are the most theoretical-
ly inclined and the ones we would expect to be associated with sociology due to the word embeddings found in its context. Adding to the interpretation of Figure 24, there seems to be an association between all centroids, but the similarity in distribution appears most strongly between the science centroid and the philosophy centroid and, thereafter, the sociology centroid and the theory centroid. By inferring the results of the previous analysis of the abstract corpus, we would assume philosophy, and to a lesser extent theory, to be words related to qualitative sociology (quantitative/qualitative), whereas science ought to be associated with quantitative sociology (quantitative/qualitative). This does not seem to be the case here but rather all these centroids look to be closely related and certain styles that perhaps would be associated with qualitative sociology are close to all of them.

Thus, the question that arises is whether there is another word embedding besides sociology that is suitable for finding the distributional meaning of sociological knowledge in the corpus. We can then ask, what is the research object of sociology that sociology dissertations ought to capture in their studies? Looking at the etymology of the French word sociologie, which gave birth to the English word sociology, it is built on the Latin word socius, which traditionally meant “companion” but came to mean “society”, and logia or λόγια is the old Greek word equivalent to the modern word “discourses” that came to mean “the study of”. So, at least in basic etymological terms, sociology ought to refer to “the study of society”. Therefore, the next natural step for finding the distributional meaning of sociological knowledge is to scrutinize the word embedding society and its closest neighbors in vector space.

Distributional Meanings of Society

In the previous section, the analysis was centered around the word embeddings sociology, which we learned was strongly associated with theory, philosophy, and science, which were all associated more strongly with particular styles in the corpus. Moving on to the overarching research object of sociology, that is society, this section will be grounded on the word embedding society, and follow the same procedure as the previous section. That is, we start by mapping out the distributional meaning of society by looking into its five closest neighbors, then the five closest neighbors to these neighbors, and, finally, their five closest neighbors. Based on which major con-

---

cepts are found in this analysis, we will look into how they are distributed over the corpus based on the categories of time, place, and linguistic style.

Figure 26. A network based on the text2vec skip-gram algorithm, here modeled on the full-text sociology dissertation corpus where the word embedding “society” is targeted, its five closest word neighbors, and their five closest word neighbors. Edge thickness indicates closeness in vector space and node size represents degree.

Figure 26 is a network representation of the actual distances between word embeddings in the vector space modeled on the sociology dissertation full-text corpus generated to map the distributional meaning of society. The five closest neighbors to society are culture, economy, state, as well as the interconnected word embeddings societies and modern, which together form the four corners of the word network. The interpretation that will be put forward here is that these corners represent four major themes constituting the main distributional meaning of society based on how it is used in the
corpus of sociology dissertations, namely by addressing aspects of culture, economy, politics, and modernity. To walk the reader through how this understanding was reached, as well as the distributional meaning of each corner of the network, we will have to look closer into its constituting parts.

Beginning with the aspect of culture, its main word embedding, culture, is connected to two embeddings sharing the same word stem, namely cultural – connecting to heritage, capital, economic, and linguistic – and cultures – leading to eras, traditions, epochs, and back to societies. While the word cultures is interpreted to refer mainly to historical aspects of societal entities on the macro level, cultural seems to address cultural heritage, language, and the forms of capital associated with the highly influential sociologist Pierre Bourdieu\textsuperscript{140} – where cultural capital is one and economic capital another. Another close word embedding to culture is language, which is also connected to linguistic as well as languages, spoken, proficiency and skills. In comparison to cultural, this word embedding is hereby interpreted to be more specifically related to the linguistic aspects of culture, with what seems to be particular attention given to different languages and the “proper” way they ought to be spoken. Further, culture is neighbored by religion, which has a context consisting of similar words – religions, religious, and religiosity – as well as the specific religion islam and its followers muslims. The last close neighbor of culture is popular, which is linked to media, television, mass, and movements. While the word mass is understood as a synonym to popular, and the word movements is thought to address social movements, media and television are construed to be mediums where culture is transmitted. By assembling all word embeddings forming the context of culture, its distributional meaning appears to be centered around general words for culture, with an emphasis on history, language, religion, and the media, as well as a specific sociological study of cultural capital. Thus, it is harder to pin down the cultural aspect of society in one or a few disciplines in comparison to the economic and political aspects. The diverse set of topics spans over several cultural sciences, such as ethnology and religion, but also social sciences like anthropology, as well as media and communication studies.

Continuing with the economic corner of the left of Figure 26, one finds that economy is linked with three types of word embeddings. We have two with the same stem, economic and economies, where the first leads to socioeconomic, resources, and financial, as well as to political and cultural, and the second to, on the one hand, societies and countries, and, on the other, markets and firms. While the distributional meaning of economic is

\textsuperscript{140} In a study of the mainstream sociology outputs, Bourdieu was found to be the most cited and influential sociologist in the world during the 2010s (Korom, 2020).
here understood to pick up both references to general economic indicators and instances where the economic is compared to the political and the cultural, it seems as though economies can both be something of a synonym to societies and countries as well as capture specific economic entities. There is also a type based on capitalist and capitalism—which is also connected to capitalisms—that forms a triangle with capitalists. The former word embedding, capitalist, leads to coloniality and exploitation and the latter, capitalism, further connects with capitalisms and socialism. Thus, it appears as though the distributional meaning of capitalism might be slightly more neutral as an economic system together with socialism, while capitalist is more often attached to words with seemingly negative connotations like coloniality and exploitation. Lastly, economy is linked to gift that is neighbor to gifts, exchange, exchanges, reciprocity, and mauss. These words are assumed to be related to the bond between the exchange-based gift economy and reciprocity as theorized by Marcel Mauss in his classic work The Gift from 1925. In summary, the distributional meaning of the economic aspect of society is interpreted to be a mix of comparison between either different economic systems, or economy vis-à-vis other realms of society, or a more direct critique of the capitalist system. These are all topics traditionally dealt with within the economic sciences.

To the north in Figure 26, we find the political aspect of society stemming from the word embedding state. On the left-hand side of state, its close neighbor political is linked to politics, parties, ideological, and, as we learnt in the previous paragraph, economic. The first set of words are understood to refer to general state politics based on political parties adhering to different ideologies. Turning to the right-hand side, state is connected to its plural form states, a word embedding in close distance to europe, countries, nations, and united. Thus, we can assume that states is a word used rather synonymously to countries and nations, where Europe appears to be a typical set of states addressed, as well as the particular country the United States that is constituted by this particular word embedding. To the north, state forms a triangle with government and administration. The former leading to governments, authorities, parliamentary, and parliament, and the latter to ministries, administrative, and administrations. The distributional meaning of this substantial set of words is interpreted as words for institutional bodies constituting the state apparatus. Lastly, state connects to welfare that is neighbored by policies, policy, redistributive, and provision. These are interpreted to refer to the welfare state that reallocates resources and services via policy-making. Adding together all word embeddings forming the context of state, its distributional meaning is interpreted to refer to political parties and ideologies, the workings of the state apparatus, and macro societal entities like nations and countries, where the welfare state appears to be given an advanced position. These political top-
ics, spanning from ideology to governance, form the pinnacles of political science.

In the last corner of the Figure 26, on the right side of the network, lies the entangled aspects of, on the one hand, societal comparisons and, on the other, modernity. This corner is based on two of the five closest neighbors to society, namely the interconnected word embeddings societies and modern. These two word embeddings are interlinked with both contemporary and western, which already indicates what sorts of society are addressed, both for societal comparisons and for addressing modernity. The context of western picks up both what is understood to be neutral relational directions, namely eastern and west, as well as a specific focus on one continent, european and europe, which is perhaps implying that particularly Western Europe is often referenced. Together with western, societies also connects to industrialized that is neighbored by industrialised, advance, and countries. As has already been touched upon above, the word embeddings societies and cultures are also connected, and countries is a word embedding that is linked to both states and economies. Thus, this part of the network is interpreted to reference societal comparisons on what is presumed to be a macro scale. By reading through the word connotations forming this wider context, we can assume that a certain type of society is in focus for these studies, namely those that are advanced, industrialized, and situated in the West or Europe. This aspect of societal comparisons is not specific to sociology but is most typical for social sciences in general.

Turning to contemporary, this not only forms triangles with western, societies, and modern, but is also connected to what are seen as other words for temporality, todays and historical, as well as sociological, which we learned in the previous section is heavily associated with theory. The key word embedding modern is further connected with premodern that leads back to societies but more specifically has a context consisting of peasant, curio, decrepit, and medieval, which are understood to be words referencing things primarily belonging to a premodern time. Further, there is a triangle consisting of modern, modernity, and late. Like contemporary, late appears to pick up words referencing time — early, era, and century — but these words seem to address temporality on a wider scale. On the other hand, modernity connects further to postmodernity, modernisation, giddens, and beck. We here find what is understood to be references to the process of modernization, to the eras of modernity, late modernity, and postmodernity, as well as to two modern sociologists famous for theorizing

141 It is hardly the case that the peasant is solely a premodern role and occupation, but the word is here understood in the context of the agrarian societies predating industrialization and the rise of modernity as a phenomenon.
about modernity and modernization, namely Anthony Giddens and Ulrich Beck. The words in this set are interpreted to be cornerstones in the fifth and last aspect of society found in the corpus, namely that of modernity. Adding on the different words for temporality, we can here assume references to the historical development or evolution of society into the modern age, which is a key aspect of modernity. Of all five aspects of society found in Figure 26, modernity is construed to be the most prototypical for sociology since it is a (if not the) problem dealt with by the founding figures of the discipline at the turn of the 20th century.

To summarize the investigation of Figure 26, the analysis of the network, which consisted of the closest word embeddings to society and its neighbors, represents the distributional meaning of society as it is expressed in the full-text dissertation corpus. The interpretation put forward is that there are five major aspects of society that the sociology dissertations scrutinized here emphasize. These are the cultural aspect of society, the economic aspect, or the political aspect, as well as one of the two more intertwined aspects, societal comparisons or modernity. To uncover what part of society is in focus for what part of the corpus, we will move on to investigate how these five aspects of society are distributed over time, place, and style. As in the previous section with the distributional meaning of sociology, we will begin by looking into the temporal dimension of the distributional meaning of society and then head on to the spatial and stylistic dimensions in tandem. As usual, the dissertations are sorted into their year of publication, university of publication, and stylometric category to represent the three dimensions investigated. Again, small dictionaries were constructed to capture the concept of interest in vector space, a so-called centroid, to be able to investigate the distance between it and the sociology dissertations. The centroids investigated, which are created based on the aspects of society mapped out in Figure 26, are culture, economy, modernity, politics, and societal comparisons.

142 The dictionary forming the culture centroid consists of the following word embeddings: culture, cultural, religion, ethnicity, media, language.
143 The dictionary forming the economy centroid consists of the following word embeddings: economy, economic, capitalism, capitalist, market, financial.
144 The dictionary forming the modernity centroid consists of the following word embeddings: modernity, modern, postmodern, premodern, modernisation, modernization.
145 The dictionary forming the politics centroid consists of the following word embeddings: politics, political, policy, state, government, welfare.
146 The dictionary forming the societal comparisons centroid consists of the following word embeddings: societies, countries, states, nations, industrialized, industrialised.
Figure 27. A time series of the distance between the dissertation full-texts and five dictionary-based centroids (culture, economy, modernity, politics, and societal comparisons) which are conceptualized to represent the distributional meaning of “society” in vector space.

Figure 27 is a visualization of the temporal distributions of centroids interpreted to be close to the word embedding ‘society’ in vector space. We can here see that the economy centroid and the societal comparisons centroid follow a similar pattern with a peak ($\approx 0.7$) at the beginning of the time period, followed by a fast drop in 2001 and a stable wobbling around the zero value to 2017, only to drop in 2018 and a rise again in 2019. The difference between the two is that the economy centroid is more prevalent in the 2000s with a rather distinct rise in 2008 ($\approx 0.3$) and a deeper slope in 2018 ($\approx -0.7$), while the societal comparisons centroid has a more stable overall presence and higher values in the 2010s. The culture centroid and the modernity centroid also express somewhat comparable patterns over time, but to a lower
degree than the previously mentioned centroids. Beginning with a drop (≈ -0.6) in 2001, the culture centroid shows a small overall positive presence throughout 2003-2015 (≈ 0.2), only to express another negative trend in 2016-2019 and return to its lowest values (≈ -0.6) for 2017-2018. Quite similarly, modernity had a slight positive trend (≈ 0.1) through the majority of the period (2000-2015), with two high points (≈ 0.5) in 2005 and 2013, but a more distinct negative trend in the last years (2016-2019), with its low-point in 2018 (≈ -0.8). The last centroid, the politics centroid, lies somewhere in between the other two groups of centroids in terms of time, with a small positive manifestation throughout the 2000s, with two peaks (≈ 0.5) around 2000 as well as 2009, and a negative manifestation in the 2010s, with an all-time low in 2018 (≈ -0.8).

A quick look at Figure 28 reveals a sometimes-accentuated internal difference between the linguistic styles and the universities, giving the first indication that style and place seem to matter in terms of what aspect of society is in focus. In comparison to the spatial university category, the style category has shorter boxes (i.e., more concentrated values between the upper and lower quartile) but a larger difference in mean values within the category. This pattern holds for all five aspects of society and, as was the case for the distributional meaning of sociology, this suggests that styles might be a better indicator than place for mapping out different aspects of the distributional meaning of society. For the linguistic styles, the clearest opposition between category members is found in the modernity centroid, and the corresponding centroid for the universities is the societal comparisons centroid. However, even for the societal comparisons centroid, the linguistic styles are still more internally diverse than the universities. Nevertheless, we should walk through each centroid to see if they reveal any interesting pattern concerning the effect place and style might have on the distributional meaning of society.
Figure 28. A comparison of how close the dissertation full-texts, divided into groups of style and place, are to the five dictionary-based centroids (culture, economy, modernity, politics, and societal comparisons) conceptualized to represent the distributional meaning of “society” in vector space.

For the university categories visualized in Figure 28, Uppsala and Gothenburg appear to have the most concentrated distribution of dissertations over the five centroids, while Lund and Umeå seem to have the least concentrated, and Stockholm is somewhere in between. This can be interpreted as the dissertations stemming from the first-mentioned universities are more prone to refer to the same aspects of society in comparison to the rest. In terms of the specific centroids, the modernity centroid is characterized by the relatively high mean value of Uppsala and low mean values of Stockholm and Gothenburg, whereas Lund and Umeå are in the middle, with wider spreads
between the upper and lower whiskers. Uppsala also has the highest mean values for the culture centroid together with Umeå, while Stockholm is at the far lower end, and the remaining two universities are centered around the zero value. For the societal comparisons centroid and the economy centroid, Stockholm and Umeå have a relatively high mean value, while Uppsala and Lund have low values. The difference is slightly more accentuated in the latter centroid, where Stockholm has a higher score than Umeå, and Uppsala lies lower than Lund. Lastly, the politics centroid is distinguished by a relatively high value for Umeå and a small, elevated value from Gothenburg, whereas the rest of the universities express a negative mean value. Taking a comparative perspective on the universities, dissertations from Umeå stand out with positive mean values in four centroids and a neutral one for the modernity centroid. A somewhat similar but more negative distribution is seen in Gothenburg, scoring close to zero in the same four centroids and a more distinctively negative score in the modernity centroid. While Uppsala has high mean values and Lund has zero mean values in the modernity centroid as well as the culture centroid, both universities score low on dissertations in the other centroids. Almost the opposite pattern to Uppsala is true for Stockholm, except for the fact that it has a low mean value in the politics centroid. Perhaps we can sense some sort of opposition led by Uppsala and its modernity and culture centroids, on the one hand, and Stockholm with its societal comparisons and economy centroids, on the other. However, this potential pattern will have to be investigated further.

Moving over to the style category and its wider internal differences, it appears that the person-oriented style is the one with the most even focus over the centroids, while the rest of the linguistic styles are characterized by sometimes having a very concentrated distribution of dissertations over the centroids and sometimes not. For the specific centroids, the modernity centroid stands out the most by having the highest and lowest mean values in Figure 28, represented by the concept-oriented style and the distribution-oriented style, respectively. In between the two styles, we find a slight positive mean value for the relation-oriented style and a modicum of negative for the person-oriented style and the circumstance-oriented style. The culture centroid is more balanced between the linguistic styles, with a slightly elevated mean value of closeness for the concept-oriented style, the relation-oriented style, and the person-oriented style, while the circumstance-oriented style has a slight negative mean value and the distribution-oriented style even more so. Societal comparisons and economy are two centroids that appear to be quite similar in their distributions within the style category, yet the differences between styles are more widespread for the latter centroid. The pattern is that the highest mean value is found in the distribution-oriented style, followed by a small negative value for the critical and the circumstance-oriented styles, as well as an even lower value for the concept-
oriented style and the circumstance-oriented style. Finally, the politics centroid is characterized by a high mean value for the relation-oriented style and low mean values for the concept-oriented style, the person-oriented style, and the distribution-oriented style.

When comparing the linguistic styles to each other, the concept-oriented style stands out with a very high score in the modernity centroid and a small positive in the culture centroid, very low in societal comparison, and mildly low in economy and politics. A similar pattern is expressed by the relation-oriented style, with a relatively high score in modernity and culture respectively, and somewhat low in societal comparisons and economy, but this style has the highest score of all styles for the politics centroid. Except for a small positive in culture, the person-oriented style is low in all other centroids. This negative pattern is even more accentuated in the circumstance-oriented style, which has no positive mean value and negative values in all centroids, where the politics centroid is the closest to zero. The distribution-oriented style, in vast contrast to the concept-oriented style and the relation-oriented style, is positive in societal comparisons and economy, and negative in the rest. Thus, it appears there is no specific style that is comparably close to the word embedding society and its neighbors, but rather that the linguistic styles contribute to different aspects of the collected distributional meaning of society. However, whether there is an opposition between the linguistic styles, as was suggested in the previous chapter, will have to be investigated in more depth.

The analysis put forward in this section has insinuated a linkage between the societal comparisons centroid and the economy centroid. This suggests that macro comparisons of society go hand in hand with analyses of the economic aspects of society, like capitalism and markets. These two centroids appeared to be most associated with the distribution-oriented style as well as the universities Stockholm and Umeå. Another relationship, albeit a comparably weaker one, also appears to be found between the modernity centroid and the culture centroid, as well as the conceptual style and the culture centroid, suggesting that themes like premordernity, modernity, and postmodernity might be related to cultural aspects of society like ethnicity, media, and language. In this case, there seems to be an association between the modernity centroid and the concept-oriented style and, to a much lesser degree, the circumstance-oriented style as well as Uppsala. For the culture centroid, the association is weaker but the strongest candidates are the concept-oriented style, the circumstance-oriented style, Umeå, and Uppsala. Thus, if there is a division in the corpus in terms of how society is approached, it lies between dissertations belonging to the distribution-oriented style and the concept-oriented style and, which is less distinct, between those defended in both Stockholm and Uppsala.
Both the analyses put forward in the previous and the present chapters, centered on sociology and society, respectively, have given us a deeper understanding of the distributional meaning of these word embeddings and hinted toward the interpretation that there is a stronger stylistic than spatial pattern of what aspects of sociological knowledge are in focus. There have also been some results pointing in the direction that different styles and, to a lesser extent, universities might perhaps be associated with different sets of concepts forming these distributional meanings. This tendency sends us back to the analytical theme initiated in the first empirical chapter and continued in the following two, namely whether sociological knowledge appears to be divided. The studies of the dissertation abstract corpus suggested proof of the so-called methodological divide. To investigate its plausibility in the word embeddings modeled on the dissertation full-text corpus, we will have to move beyond the word embeddings of sociology and society. This is because the study of sociology led to the interpretation that a specific set of dissertations associated with a certain linguistic style are the ones that tend to address the word sociology. Further, the study of society exposed five central aspects of society seemingly not derived from methodology.\(^{147}\) Instead, the most alluring strategy will be to return to the sociological dichotomies spelled out in the theory chapter, and map out their distributional meanings in vector space as well as their associations with the places occupied by different categories of dissertations.

**Distributional Meaning of Dichotomies**

The analyses of positions of the word embeddings sociology and society in the vector space modeled on the full-text dissertation corpus led to the query of whether there are any dividing lines in the distributional meaning of sociological knowledge, as there seemed to be in the dissertation abstract corpus. To investigate this question further, we will have to return to some of the prevailing dichotomies proposed by theoreticians to divide sociological knowledge that were spelled out in the theory chapter. The word embedding forms of the dichotomies initially investigated are quantitative/qualitative, empirical/theoretical, macro/micro, objectivism/subjectivism, positivist/interpretivist, and realism/constructivism. This selection is made to enable a deeper exploration of the dichotomies suggested to divide the word constellations and thematic arrangements of the abstract corpus as well as the linguistic styles of the full-text corpus. To get a glimpse of the context

\(^{147}\) Perhaps the societal comparisons centroid is somewhat of an exception to this rule since it appears to pick up words for large-scale comparison that are associated with quantitative methodology, but one could also imagine, for instance, qualitative historical studies that are societal comparisons on the macro level.
for the poles of each dichotomy, their five closest neighbors will first be investigated in the representation of a word network. Based on this investigation, five different dichotomies will be singled out to form the basis of analyses looking into the temporal, spatial, and stylistic distribution of each dichotomy.

Figure 29. A network based on the text2vec skip-gram algorithm and modeled on the full-text corpus, showing the closeness between different word embeddings in vector space. The network is based on the five closest neighbors of 12 word embeddings that carry the same names as contrast words forming six sociological dichotomies (which were explained in the theory chapter). The thickness of the edges indicates how close the word embeddings are to each other in vector space, and the size of the nodes is constant.

Figure 29 is a network based on the five closest neighbors to the 12 word embeddings distributed over the six dichotomies addressed above. The word embeddings are represented by blue nodes of the same size and the edges indicate the distance between the neighboring words in vector space. In the figure, we find three small networks based on one dichotomy each and one larger network constituted by three dichotomies and their neighbors. Thus, three of the dichotomies follow the parallel pattern where the word embed-
Thus, the three mentioned dichotomies are interpreted to discuss different aspects of methodology and the discussions seem to mostly stay within the dichotomy. We could imagine that qualitative and micro would form one network and quantitative and macro another, which would strengthen the interpretation that key terms within one side of the methodological divide share a similar distributional meaning. This seems, however, not to be the case. Rather the distributional meaning of, for instance, micro appears to be almost identical to macro. The reason for this is that the words constituting the dichotomies are discussed in the same context, that is have the same specific neighboring words in vector space. In other words, there are presumably a good share of sentences where micro is contrasted with macro and meso. This network is also interpreted to be the most specific for social science since it is believed to relate to different analytical levels of society. The most divergent of the three dichotomies is empirical/theoretical – which is constituted by the most generic words in the sense that they are used within a good share of the sciences and the humanities – but even here there is a strong connection between the word embedding representations of the poles, empirical and theoretical, forming a triangle with analytical. In between these two dichotomies, we find quantitative/qualitative, comprising methodological words commonly used in the social sciences, but the word embedding qualitative would presumably not appear in a dissertation dedicated to one of the ’hard’ sciences.
The three last dichotomies are not only interconnected internally — objectivism with subjectivism, realism with constructivism, and positivism with interpretivist — but also externally to each other through relativism, constructivism, as well as positivism. The word embedding objectivism is also linked with hermeneutics and shares two neighbors with subjectivism — the mentioned constructivism as well as structuralism. In turn, subjectivism is further connected to subjectivist and atomism. By only looking at the first of the two dichotomies, it seems evident that sometimes seemingly incommensurable positions, rather than what have traditionally been seen as commensurable positions, share the same context (e.g., objectivism and hermeneutics, or subjectivism and atomism). The second dichotomy, constituted by realism and constructivism, are interlinked and connected to relativism. As was mentioned, realism further connects to positivism but also to realist and ontology, while constructivism — the word embedding neighbor to both objectivism and subjectivism — is further linked to constructivist, constructionism, and constructionist. There are few surprising links in this dichotomy, since all words with the stems “construct” or “real” are firmly interrelated, and relativism relates to the discussion of constructivism versus realism. The last dichotomy, positivist versus interpretivist, connect to each other, and while the former links to the rest of the network through positivisms, the latter does so by relativism. Further, positivist is connected to scientistic, positivist, and science, whereas interpretivist is connected to cognitivism, antipositivist, holist, and theoreticism. Here we find several key concepts and occasionally prejudices in the debate between positivism and interpretivism, such as that the interpretivist position is an “antipositivist” response to scientistic positivism by drawing on holism, but also one that can be accused of relativism, as well as cognitivism and theoreticism.

The overall interpretation of the three last dichotomies put forward is that they signify ontological and epistemological views that are discussed in the same context rather than two different contexts. That is interpretivist, antipositivist, subjectivism, cognitivism, constructivism/constructionism, holist, and hermeneutics, on the one hand, and their respective “antithesis” scientistic, positivist, realism, objectivism, atomism, and positivism, on the other hand. These intertwined philosophical positions are commonly referred to within sociology but go far beyond the social sciences in their scope and heritage. Due to their interconnectedness in Figure 29, they will form a single dichotomy in the forthcoming analyses conducted within this section.

If summarizing the interpretation of Figure 29 put forward in the paragraphs above, we have found that three of the dichotomies are isolated and to dif-
ferent extents pick up the same neighbors, whereas the other three are interconnected. Thus, we find ourselves with four major dichotomies for investigating the distance between them and the sociology dissertations in vector space in terms of time, place, and style. Due to the fact that the poles of all dichotomies are each other’s closest neighbor, it is clear that it is a bad idea to formulate dictionary-based centroids only from these word embeddings. For instance, since macro and micro are two neighboring word embeddings in vector space, investigating the sociology dissertations’ closeness to the context of micro and macro, respectively, would probably cancel each other out rather than signify whether the dissertations are primarily studies of the micro level or the macro level. Similarly, qualitative and quantitative, empirical and theoretical, subjective and objective, positivist and interpretivist, share the same context. Thus, to serve the purpose of exploring the distribution of the dichotomies within the full-text corpus, we have to generate wider dictionaries that include more words associated with the oppositions captured within each dichotomy. That makes up a total of eight dictionaries, one for each pole of each dichotomy: empirical\textsuperscript{148}/theoretical,\textsuperscript{149} macro\textsuperscript{150}/micro,\textsuperscript{151} positivist\textsuperscript{152}/interpretivist,\textsuperscript{153} and quan-

\textsuperscript{148} The dictionary forming the empirical pole of the empirical/theoretical dichotomy consists of the following word embeddings: empirical, data, empirics, empirically, observation, observations, experiment, experiments, experience, experiences, informant, informants, respondent, respondents.

\textsuperscript{149} The dictionary forming the theoretical pole of the empirical/theoretical dichotomy consists of the following word embeddings: theoretical, theory, theories, conceptual, concept, concepts, principle, principles, proposition, propositions, theorist, theorists, theoretician, theorists.

\textsuperscript{150} The dictionary forming the macro pole of the macro/micro dichotomy consists of the following word embeddings: macro, country, countries, people, state, states, population, populations, structure, structures, structural, discourse, discourses.

\textsuperscript{151} The dictionary forming the micro pole of the macro/micro dichotomy consists of the following word embeddings: micro, action, actions, interact, interacting, interaction, interactions, conversation, conversations, conversational, interpersonal, talk, talks.

\textsuperscript{152} The dictionary forming the positivist pole of the positivist/interpretivist dichotomy consists of the following word embeddings: positivism, positivist, objectivism, realism, objective, cause, causes, causal, causation, causality, model, models, mechanism, mechanisms, effect, effects, factor, factors, scientific, hypothesis, hypotheses, hypothetical.

\textsuperscript{153} The dictionary forming the interpretivist pole of the positivist/interpretivist dichotomy consists of the following word embeddings: constructionism, interpretivist, subjectivism, constructivism, subjective, meaning, meanings, meaningful, meaningmaker, meaningmaking, understanding, understandings, interpretation, interpretations, construct, construction, narrative, narratives, humanistic, intention, intentions, intentional.
The dictionary forming the *quantitative* pole of the *quantitative/qualitative dichotomy* consists of the following word embeddings: quantitative, survey, surveys, questionnaire, questionnaires, statistics, statistical, dataset, datasets, cohort, cohorts, census, censuses, panel, registers, database, databases.

154 The dictionary forming the *quantitative* pole of the *quantitative/qualitative dichotomy* consists of the following word embeddings: qualitative, interview, interviews, recording, recordings, ethnography, ethnographic, fieldnote, fieldnotes, transcription, transcripts, archive, archives, fieldwork, fieldworks, document, documents.

155 The dictionary forming the *qualitative* pole of the *quantitative/qualitative dichotomy* consists of the following word embeddings: qualitative, interview, interviews, recording, recordings, ethnography, ethnographic, fieldnote, fieldnotes, transcription, transcripts, archive, archives, fieldwork, fieldworks, document, documents.
tivist/interpretivist, and quantitative/qualitative. A line in the time series reflects the mean distance of all dissertations produced that year (where 0 indicates the average of the whole corpus) to two fixed sets of word embeddings. The algorithm is set to treat these two sets of word embeddings as opposing poles where one set represents negative values and the other set positive values. In Figure 30, each line represents a sociological dichotomy where closeness to one pole result in a positive value and closeness to the other in a negative value. For instance, the mean of the dissertations defended in 2005 is closer to the word embeddings picked up in the dictionary of the qualitative pole than that of the quantitative pole and, thus, they are given a negative value one the line representing that dichotomy in the time series.

The first thing to notice about Figure 30 is that the values are slightly more centered around the horizontal zero line than in the previous time series (i.e., Figure 24 and Figure 27). This suggests that the word embeddings forming the poles of the dichotomies have a fairly equal distribution over time in vector space. When comparing the different dichotomies, we find an almost identical distribution over time for word embeddings within the quantitative/qualitative dichotomy and the positivist/interpretivist dichotomy. That is, when the word embeddings meant to signify words associated with the quantitative pole go up so do those found within the positivist dictionary, and when word embeddings representing the qualitative pole are addressed so are those related to the pole interpretivist. Both dichotomies are most influential at the very start and very end of the time period but less so around 2008-2016. To a lesser extent, the trend of these two dichotomies resonates with the macro/micro dichotomy, but this dichotomy is stronger throughout the 2000s and weaker in the 2010s. In contrast to word embeddings forming the macro/micro dichotomy, those of the empirical/theoretical dichotomy seem to follow an opposite trend and are underperforming somewhat in the 2000s but sturdier in the 2010s. In other words, closeness to word embeddings of the empirical pole and micro pole is related, or, to put it differently, word embeddings encapsulated in the theoretical dictionary increased with word embeddings in the macro dictionary.

The main take-home message from Figure 30 appears to be that word embeddings adhering to all respective dichotomies occur rather consistently over time, that primarily word embeddings of the positivist/interpretivist dichotomy and quantitative/qualitative dichotomy are most strongly associated, as well as those of the empirical/theoretical dichotomy and the macro/micro dichotomy, which follow almost inverted patterns over time. On this basis, we can head on to investigate how the four dichotomies are distributed over place and style. This analysis is based on Figure 31, which
shows how close the sociology dissertations, sorted into styles and universities, are to the word embeddings constituting each pole of four dichotomies.

Figure 31. A comparison of the distance between the sociology dissertations, divided into groups of style and place, and the eight poles conceptualized to represent the distributional meaning of four different sociological dichotomies (quantitative/qualitative, positivist/interpretivist, macro/micro, and empirical/theoretical) in vector space.

As was the case in the previous sections, the distribution found in Figure 31, where the mean values of the linguistic styles are more diverse from each other than those of the universities, seems to suggest that the linguistic styles
are more distinctive than the universities. In this case, the linguistic styles also appear to show less dispersed selections of dissertations than the universities and to be more concentrated around the mean values. The clearest examples are the quantitative/qualitative dichotomy and the positivist/interpretivist dichotomy where the sociology dissertations in each style are tightly clustered together, something that is not the case for the university category. To get a better idea of how the sub-corpora representing linguistic style and locality relate to each other, we have to first inspect them separately.

All in all, the universities’ closeness to either side of the dichotomies follows a pattern we are familiar with from the previous empirical chapters. This pattern begins with dissertations defended in Stockholm at one end, through those coming from Umeå and Gothenburg, and finishing with those stemming from Lund and Uppsala at the other end. The exception is the empirical/theoretical dichotomy where Gothenburg scores as high as Stockholm, and both are followed by Umeå and Lund and then Uppsala. Therefore, a plausible interpretation might be that Stockholm and, much less so, Umeå are the universities most inclined to sociological knowledge related to the quantitative, positivist, macro, and (together with Gothenburg this time) empirical poles. In contrast, dissertations defended in Uppsala and Lund have produced sociological knowledge that is more qualitative, interpretivist, micro, and (mostly for Uppsala) theoretical in its content. However, the evidence here is not all that convincing since the universities are rather close to each other, and for each university there are sociology dissertations that are on both extreme ends of the mean value.

The distribution-oriented style has a suggestive top score followed by neutral or slightly negative mean values for the relation-oriented style and the circumstance-oriented style and ending with even slightly more negative scores for the person-oriented style and the concept-oriented style. Again, the empirical/theoretical dichotomy makes an exception with a positive score from the distribution-oriented style, the person-oriented style, and the circumstance-oriented style, whereas the relation-oriented style is slightly negative and the concept-oriented style takes a clear negative position. Thus, it seems like the distribution-oriented style is most clearly associated with sociological knowledge that is quantitative, positivist, and on the macro level. The opposite position – that is, qualitative, interpretivist, and micro – does not have a corresponding candidate to the distribution-oriented style; rather all other styles appear to tilt towards it. Among these four styles, the person-oriented style and, even more so, the concept-oriented style, appear to be the ones most associated with this type of sociological knowledge. As for the linguistic styles’ closeness to empirically or theoretically laden word embeddings, the concept-oriented style appears to take the uttermost distinct
position at the theoretical pole, followed in the distance by the relation-oriented style, while the other styles seem more inclined to the empirical pole. While three of the linguistic styles end up in rather close adjacency to each other in at least one of the dichotomies, it seems safe to say that the concept-oriented style and the distribution-oriented style represent rather antithetical positions in the distributional meanings of sociological knowledge as it is conceptualized through the sociological dichotomies.

Closing Thoughts on Distributional Meanings

The focus of this chapter has been a vector space comprised of word embeddings generated with the text2vec skip-gram algorithm (cf. Mikolov et al., 2013) which were analyzed by interpreting the distributional meanings of how concepts suggested to articulate sociological knowledge are used in the full-text corpus. By adding the method of Concept Mover’s Distance (Stoltz and Taylor, 2019), it was possible to investigate the distance of the full-texts in relation to all word embeddings. If one full-text is close to a set of word embeddings and another full-text far away, this means that the former “taps in” to these word embeddings (in terms of all the words this text consists of in relation to the vector space that is generated from all full-texts) and the latter does not. Building on the strategy leading the previous analyses, the full-text corpus was sorted into groupings based on the years the dissertations were defended (2000-2019), the universities where the defense took place (Gothenburg, Lund, Stockholm, Umeå, and Uppsala) as well as the five styles that were generated in the previous chapter (the circumstance-oriented style, the concept-oriented style, the distribution-oriented style, the person-oriented style, and the relation-oriented style).

The analyses presented here have proposed that the distributional meanings of sociological knowledge can be operationalized as the contexts surrounding the word embedding of ‘sociology’, the word embedding of ‘society’ (with sociology conceived as the science of society), and word embeddings articulating the sociological dichotomies laid out in the theory chapter. Through a network visualization of the closest neighbors to the word embedding ‘sociology, it was suggested that their context primarily consisted of words related to theory, science, and philosophy. Turning to the network based on the word embedding ‘society’, its most adjacent neighbors were interpreted to address the economy, politics, culture, modernity, or societal comparisons.

Following the interpretation of the word embedding ‘sociology’, it was proposed that we should not just look at how close the full-texts were to words
for sociology but also for theory, science, and philosophy. Thus, for each group of words, a dictionary was constructed that could be utilized to compute a centroid holding the average distance of the full-texts to these words. Looking at the aspect of time (the years the dissertations were defended), the four centroids were suggested to follow almost the same fluctuating pattern with a drop at the beginning and end of the time period but otherwise a rather consistent overall closeness to the corpus throughout the time period (2000-2010). This seemed most clear for the sociology centroid and the theory centroid, while the philosophy centroid and the science centroid look like they more often deviated from the trend of the other centroids, with the first being slightly more associated with the 2000s and the latter with the 2010s. The general interpretation proposed here is that all these centroids go together, which is an indication of that the same documents ‘tap into’ them.

Turning to the aspect of place, it was possible to see a small separation of university groupings of full-texts between the relatively positive Uppsala University and relatively negative Stockholm University and Gothenburg University. However, it was hard to find a general division between all universities since, for instance, full-texts dissertations defended at Lund University and Umeå University scored quite similarly. This result suggests that not one university grouping of full-texts stands out drastically. In other words, like time, place does not appear to be a major factor for who is forming the distributional meaning of ‘sociology’ in this corpus. Concerning the previous chapters that suggested a comparatively clear spatial division between dissertations from Uppsala and Lund on one side and those deriving from Stockholm and Umeå on the other, the spatial result is somewhat unexpected, and is therefore one to which we will return in the concluding discussion presented in the next chapter.

When considering the categories of linguistic style, it was suggested that the concept-oriented style and, to a much lesser extent, the relation-oriented style housed the biggest share of full-texts most closely related to all four centroids (sociology, theory, science, and philosophy). In contrast, full-texts of the distribution-oriented style scored low for all centroids, particularly the philosophy centroid but, more surprisingly, also had relatively low scores in the science centroid. The tendency that dissertations categorized within the tiny concept-oriented style, as well as some dissertations falling within the relation-oriented style, are such high scorers in all four centroids alluded to the interpretation that this set of full-texts represent the main distributional meaning of the word embedding ‘sociology’ in the full-text corpus. This supports the understanding put forth in the previous analytical chapter based on stylometry (i.e., that the concept-oriented style is the linguistic style that explicitly deals with questions of sociology as an entity in itself).
Continuing with the word embedding of ‘society’ and its context, five dictionaries were created and five corresponding centroids computed – the culture centroid, the economy centroid, the modernity centroid, the politics centroid, and societal comparisons centroid – in the same manner as for the word embedding of ‘sociology’. Regarding time, it was shown that the economy centroid and the societal comparisons centroid follow a similar fluctuating pattern, with a peak at the beginning and end of the time period. Similarly, the culture centroid and the modernity centroid also express somewhat comparable patterns over time, with an overall positive presence throughout but low values at the beginning and end of the period. The last centroid, the politics centroid, was found in between the other two groups of centroids in terms of time, with a small positive manifestation.

Further, there were internal differences between the full-texts when categorized by the five linguistic styles but this time by the five universities as well, indicating that linguistic style and place seem to matter in terms of what aspect of society is in focus. When the full-texts were sorted into groups of universities, the most prominent distinction could be seen between sociological knowledge presented in dissertations defended in Uppsala and Stockholm, since when one group of full-texts scored high, the other groups scored low. Dissertations from the former university were associated with the modernity centroid and the culture centroid, and those from the latter with the societal comparisons centroid and the economy centroid. However, both groups scored low in the politics centroid, where dissertations from Umeå stood out as the only group of full-texts with high scores. For the linguistic styles, the clearest opposition was found in the modernity centroid and the culture centroid, on the one hand, and the societal comparisons centroid and the economy centroid, on the other. The former pair of centroids were closest to full-texts of the concept-oriented style and the latter to full-texts of the distribution-oriented style. Compared with each other, the linguistic style category appeared to show more concentrated values and larger differences in mean values than the spatial university category, and this pattern held for all five centroids modeled from the word embedding of ‘society’. The results suggest that at least for this corpus, linguistic style might be a better indicator than place for mapping out different aspects of the distributional meaning of ‘society’, as was the case for the distributional meaning of ‘sociology’.

Thus, in relation to the first research question focusing on fragmentation, it seemed plausible to say that in this full-text corpus of dissertations in sociology, the distributional meanings of sociological knowledge seem to be durable over time. When trying to discern what type of full-texts contribute the most to the different distributional meanings of sociological knowledge, the linguistic styles seem to be a better indicator than the universities. In answer-
ing what the computational text analyses suggest about what this sociological knowledge is all about, we have acquired an indication that the distributional meanings of sociology are mainly related to theoretical matters, whereas the distributional meanings of society are founded on different branches of empirical studies.

The analyses can be enhanced in the light of the sociological dichotomies discussed in the theory chapter, which are guiding the second research question exploring the existence of paradigms. When looking deeper into the distributional meanings of the centroids generated from the immediate context of the word embedding of ‘sociology’ the leading interpretation put forth was that the usage of words for sociology appeared to be almost synonymous with addressing theoretical concerns \((\text{empirical/theoretical})\). It was speculated that this could be one of the reasons behind the shared trend in prevalence over time between the \textit{sociology centroid} and the \textit{theory centroid}. Further, it was suggested that sociology was neighbored by two types of words, those drawing on a scholarly tradition associated with the humanities \((\text{science/humanities})\) and those closer to empirical research and the sciences \((\text{science/humanities})\). While the former was found in the adjacency of words related to philosophy, the latter was in the context of words for other social sciences like economics and psychology. However, the \textit{philosophy centroid} and the \textit{science centroid} seemed to share a similar pattern of fluctuation, which led to the idea that perhaps the usage of science words and philosophy words were interlinked in the full-text corpus but differed in the sense that science gains popularity vis-à-vis philosophy in the late 2010s. This interpretation was sustained by the low scores in the \textit{science centroid} by the \textit{distribution-oriented style} and high scores by the \textit{concept-oriented style}. The distributional meanings of the word embedding of ‘society’, however, were harder to sort into poles of dichotomies, with the exception of the dichotomy of \textit{macro/micro}. This finding is not surprising since society is a word often described as the \textit{macro} pole, while the individual represents the \textit{micro} pole.

To further study whether the sociological dichotomies were present in the full-text corpus and how they were expressed in the vector space generated by the word embeddings modeled on it, the distributional meanings of some of the key dichotomies were investigated – \textit{quantitative/qualitative, empirical/theoretical, macro/micro, objectivism/subjectivism, positivist/interpretivist, and realism/constructivism}. The results of this investigation suggested that the neighboring word embeddings for three of the dichotomies – \textit{quantitative/qualitative, empirical/theoretical, and macro/micro} – constituted isolated networks in vector space, where the overall distributional meaning was characterized by the poles referencing each other. In contrast, the three other dichotomies – \textit{objectivism/subjectivism, positivist/interpretivist, and realism/constructivism} – formed a single network con-
stituted by relations between the same type of word embeddings, but again the poles were each other’s closest neighbors.

Due to the closeness in vector space between the word embedding representations of the poles of all analyzed dichotomies, it was clear that it would be a bad idea to formulate dictionaries only from these word embeddings (see the analysis for a thorough explanation). Thus, to serve the purpose of exploring the distribution of the dichotomies within the full-text corpus, wider dictionaries of words associated with the poles of the three isolated dichotomies – empirical/theoretical, macro/micro, and quantitative/qualitative – and a single master dichotomy for the interconnected dichotomies – positivist/interpretivist – were constructed and computed to represent opposing poles in vector space via the Concept Mover’s Distance. With the scores of these eight poles for the full-text corpus, it was possible to investigate how close they were within the vector space to full-texts grouped by time, place, and linguistic style. In the case of time, the overall result was that these four dichotomies appear to occur rather consistently over the period, and that word embeddings constituting the opposing poles of the positivist/interpretivist dichotomy and quantitative/qualitative dichotomy were associated, as well as those of the empirical/theoretical dichotomy and the macro/micro dichotomy.

Turning to the third research question investigated in this dissertation, dealing with the social conditioning of sociological knowledge. The main interpretation launched in this chapter is that the linguistic styles were more distinctly separated and had more concentrated values than the universities. The clearest examples were the quantitative/qualitative dichotomy and the positivist/interpretivist dichotomy where the sociology dissertations in each linguistic style are tightly clustered together, which is not the case for the university category. Thus, the evidence for the methodological separation was not that convincing since the universities appeared to be much closer to each other in these manipulated versions of the vector space in comparison to the linguistic styles. However, all in all, the universities’ closeness to either side of the dichotomies followed the pattern we have become familiar with from the previous empirical chapters, that begin with full-texts from Stockholm at one end, through full-texts from Umeå and Gothenburg, and finish with full-texts from Lund and Uppsala at the other end. A plausible interpretation, but one that should be made with great caution due to the concentrated scores, would be that dissertations from Stockholm and, much less so, those from Umeå are most inclined to sociological knowledge related to the quantitative, positivist, macro, and (together with Gothenburg this time) empirical poles. In contrast, dissertations from Uppsala and Lund seem to present sociological knowledge that is more qualitative, interpretivist, micro, and (mostly for Uppsala) theoretical in its content.
Continuing with linguistic style, full-texts within the distribution-oriented style were shown to be most clearly associated with sociological knowledge that is quantitative, positivist, and on the macro level. The opposite position – that is, qualitative, interpretivist, and micro – does not have a corresponding candidate in the style groupings of full-texts to the distribution-oriented style; rather all other styles appear to tilt towards it. Among these four styles, the person-oriented style and, even more so, the concept-oriented style, appeared to be the ones most associated with this type of sociological knowledge. As for the linguistic styles’ closeness to empirically or theoretically laden word embeddings, the concept-oriented style appeared to take the uttermost distinct position at the theoretical pole followed at a distance by the relation-oriented style, while the other styles seem more inclined to the empirical pole. While the sociological dichotomies proved to be useful for delineating oppositions in the vector space modeled on the full-text corpus, they seemed primarily harmonized with the idea of adding the dimension of empirical/theoretical to the master dichotomy of quantitative/qualitative suggested in the previous chapter, Chapter 8, based on the same full-text corpus. This leads us to the question hinted at, at the end of that chapter, namely whether there is congruence between what the analyses of abstracts suggested and what the full-text inquiry showed?

To recall, the analyses of the abstract corpus suggested a rather one-sided picture of sociological knowledge as a dual-paradigm science based on a quantitative community and a qualitative community that followed a spatial logic, where abstracts from Stockholm and Umeå represented the former pole and abstracts from Uppsala and Lund the latter pole. However, in terms of place, the same opposition that was suggested in the abstract corpus was not as easily found in the full-text corpus. To take one example from the analysis of this chapter, full-texts stemming from Uppsala were found to be closer than average to the culture centroid and the modernity centroid, but further away from the politics centroid, the economy centroid, and the societal comparisons centroid. Yet, while full-texts from Lund shared a similar distance to the three latter centroids, it was not defined by a positive value but a zero value for the former two. Similarly, while full-texts from Stockholm had a relatively high average for the economy centroid as well as the societal comparisons centroid and a low average for the other three centroids, full-texts from Umeå were closer than average to all centroids except the modernity centroid.

Compared to the abstract corpus, then, the most striking and pervading finding of the full-text corpus seemed not to be that there is a clear divide in the corpus between two paradigms represented by two linguistic styles, such as the distribution-oriented style versus the concept-oriented style or the distribution-oriented style versus the person-oriented style. Rather, in both cases
the three remaining linguistic styles appear to stabilize the two ‘extremes’. In other words, through the computational text analyses performed on the full-text corpus, sociological knowledge does not seem to be a choice of two exclusive camps but rather that there is room for a plurality of linguistic styles. In fact, the majority of the dissertations do not appear to fall into the most specific styles such as the distribution-oriented style and the concept-oriented style, but rather in the linguistic styles that fall somewhere in between. It was further suggested that the linguistic styles could be separated not only by methodological dichotomies but also by epistemological and ontological ones. This seems to resonate more with the picture of a multi-paradigm science than that of a dual-paradigm science. This former picture is also reflected in the analyses of word embeddings, where it was suggested that the linguistic styles were associated with writing about different aspects of society. Further, the manipulation of the vector space, which was performed by telling it that words related to poles of some of the prevailing dichotomies should be oppositions, made it possible to delineate the most extreme positions that the full-texts could take and suggested that the norm of all full-texts do not follow a single master dichotomy. In contrast, the combined results of all full-text analyses resonate most strongly with a plurality of possible versions of sociological knowledge, which leads us to the aim of this dissertation, which will be thoroughly discussed in the concluding chapter.
10 Concluding Discussion

This dissertation aimed to shed new light on the crisis of sociology by empirically scrutinizing prevailing disciplinary understandings of the state of its knowledge production through computational text analysis. As explained in Chapter 1, this aim was inspired by recurring predicaments found in mainstream disciplinary outlets repeating for at least half a century that sociology is on the brink of disaster. Regardless of whether the notion of ‘crisis’ is seen as an appropriate concept for defining the state of the discipline or not, the experience of a crisis still seems to dominate the imagination of many sociologists. Due to its seemingly enduring character, the crisis of sociology could be regarded as something close to the Durkheimian ‘collective representation’. Indeed, a considerable body of literature bases itself on meta-reflections of how the discipline ought to handle the crisis within the sub-field of the sociology of knowledge, known as the sociology of sociology. While sociologists within the literature diverge in their interpretations of the evolution of sociology and what future path it ought to take, they share a common reference in the paradigm theory of science presented in the work of Thomas S. Kuhn (1962). Inspired by the Kuhnian terminology, the understanding of sociology presented in the 20th century seemed to diverge in the questions of whether sociology lost its paradigm or never had one in the first place, and whether sociology is characterized by multiple competing paradigms or is in an ‘immature’ stage of development (i.e., compared to the idea of the ‘normal sciences’ as conceptualized by Kuhn).

The 21st century brought about an increasing availability of digitized empirical material and recent developments in computational methods for analyzing text. Thus, new opportunities have emerged for conducting sociology of sociology research. After reviewing these recent trends in the field, presented in Chapter 2, three major theoretical models for conceptualizing sociological knowledge were found: fragmentation, paradigm, and social conditioning. These models laid the foundations for the research questions guiding this dissertation, presented in Chapter 5. One of the most common figures of thought is that sociology is currently fragmented, without a clear disciplinary ‘core’. However, how this fragmentation is understood is somewhat unclear in the literature. One strand of studies connects fragmentation to the fractal theory of Andrew Abbott, where sociological knowledge evolves by con-
stantly reinventing key ideas and reestablishing common forms of theoretical and methodological oppositions. An alternative strand follows the lens of paradigm theory, where sociology is perceived to lack a unifying paradigm and reside in a so-called preparadigmatic state. In both cases, sociology is perceived as a struggle between two opposing paradigms or traditions, particularly conceived within the frame of a master dichotomy separating qualitative sociology and quantitative sociology. A partly separated strand stresses the social conditioning of knowledge coined in the now classical work of Karl Mannheim by emphasizing the context of sociological knowledge production. To reiterate, it is these three key concepts suggested from the literature review that inform the three overarching research questions pointing out the direction of this empirical study focusing on the state of sociological knowledge: fragmentation (research question 1), paradigm (research question 2), and social conditioning (research question 3).

To investigate the existence of paradigms in sociology, an operationalization of paradigm theory was introduced in Chapter 3. This reading of the theory leans on its reception in the early sociology of sociology, and builds heavily on the interpretation and implementation of the paradigm concept proposed by Margaret Masterman (1970, 2009), a philosopher and pioneering computational linguist. In this implication, a paradigm can be said to exist in the discipline if a textual representation of sociological knowledge is shown to repeatedly present crude replicas of sociology. These replicas should be based on crude analogies – defined as “an analogy-drawing sequence of word-uses in natural language” (Masterman, 1970, p. 79) – which is accomplished through exercising analogical pile-up, referred to as establishing “an analogy cluster” by “simply drawing on more synonyms” associated with the crude replicas (Masterman, 2005, p. 292).

The offered theoretical strategy for delineating crude replicas of paradigms in the corpus was to establish a system of potentially recurring notions in sociology onto which sociologists pile up their crude analogies. It was suggested that throughout the history of sociological thought, the intradisciplinary discussion has tended to conceptualize sociology through a set of sociological dichotomies, nine of which, sorted into three categories, were selected. The first category, ontology, consists of three major and interrelated on-

---

156 Research question one reads: *If fragmentation is understood as a historical process where the content of the knowledge produced in a discipline becomes increasingly disintegrated, do the corpora offer evidence that sociological knowledge is in a state of fragmentation?*

157 Research question two reads: *Given the definition that a paradigm denotes a scientific community sharing a set of recurring crude replicas, do the corpora offer evidence that there are one, two, multiple, or no paradigms in sociology?*

158 Research question three reads: *Following the notion that the social conditioning of knowledge is expressed in the time and place of its production, do the corpora offer evidence that sociological knowledge is socially conditioned?*
tological dichotomies – holism/individualism, structure/agency, and micro/macro – which are often conceived as interrelated ways that help us understand how societies and individuals are constituted through opposing poles. Further, the second category is represented by three key dichotomies found to capture long-lasting epistemological queries within the discipline – objectivism/subjectivism, fact/value, and theoretical/empirical. While the six dichotomies within the ontological and the epistemological categories are all understood to prevail within sociology’s understanding of itself, the third category, comprised of three methodological dichotomies – science/humanities, explanatory/interpretative, and quantitative/qualitative – was argued to articulate some of the most principal problems contemporary sociologists have expressed that the discipline of sociology is facing.

As was proposed in the literature review of Chapter 2, this dissertation has regarded PhD dissertations in sociology, which lead to the highest educational degree in an academic discipline, as an optimal case for investigating and reviewing the validity of prevailing understandings behind the experienced crisis of sociology. Thus, this study was launched by generating a bibliography of all dissertations in sociology defended in Sweden, as presented in Chapter 4 on methodology (see also Appendix B). The reason behind the choice of dissertations is that they are all published and subjected to the same type of scrutiny by peers in this country. Based on the bibliography, the dissertations were digitized to construct the empirical material of this dissertation. The research questions, given in Chapter 5, were investigated in four empirical studies, each comprised of an analytical chapter – Chapters 6, 7, 8, and 9 – that are founded on two different corpora: 815 abstracts (1980-2019) and 380 full-texts (2000-2019). These chapters build on the use of four main types of methods that are presented in the introduction of each chapter: network analysis and cluster analysis based on word correlations, structural topic modeling (see Appendix C), stylometry based on bootstrapped consensus trees and close reading (see Appendix D), and word embeddings (text2vec, skip-gram) analyzed with network analysis and Concept Mover’s Distance (see Appendix E).

In this concluding chapter, Chapter 10, a discussion is initiated to unravel the aim of the dissertation. The chapter begins with three sections forming a discussion of the potential answers the empirical analyses can provide to the research questions posed. The first section is dedicated to presenting how the analyses of the corpora suggested that sociological knowledge is specialized but not fragmented. This position is reviewed in comparison to the results of previous studies which, in simplified terms, have tended to presume fragmentation and, at times, been explained by the fractal theory of sociology (cf. Abbott, 2001). Attending to a related issue, the second section deals explicitly with paradigms and the problem posed by previous studies utiliz-
ing the theory. In broad terms, sociology has been argued to lack a paradigm and is often described as a struggle between two opposing methodological camps – quantitative and qualitative sociology. Turning to the results of this study, the corpora of dissertations in sociology seem partly to support the existence of a methodological divide in how dissertations are presented in abstracts. However, the full-text versions of the dissertations – the actual representations of sociological knowledge in this case – seem to better reflect a multi-paradigm science with a set of paradigms focusing on different ontological and epistemological aspects of individuals and societies.

In the third section, the focus is directed towards the problem of the social conditioning of sociological knowledge presented in the literature on various temporal and spatial levels. The presented analyses suggest that both the methodological divide and the five paradigmatic candidates of sociology are conditioned by locality, expressed in the different preferences of the departments where the dissertations are defended. In this sense, sociological knowledge, as expressed in the corpora, can be illustrated as theoretical opposites embedded on the local level, forming an equilibrium on the national level. When comparing the characteristics of these corpora with those of previous studies, similarities are found that pose the question of whether diversity within international sociology is decreasing, what is referred to in the literature as institutional isomorphism (cf. Heilbron, 2014).

Building on these insights and the more experimental analyses initiated in this dissertation, the fourth section attempts to unearth the root of the crisis by suggesting that the problem lies in divergent meanings of the fundamental image of sociology’s subject matter. The solution proposed, then, is not so much to succumb to either a single paradigm or a state of ‘everything goes’, but rather to strive towards forming a flexible totality. This includes embracing the form of paradigm pluralism that comes with having a few strong paradigms able to ‘balance out’ each other. Indeed, paradigm pluralism could lead to increasing reflexivity by functioning as a safety mechanism for detecting blind spots when the explanations provided by one paradigm that are about to become too hegemonic. Lastly, the dissertation ends with a section dedicated to considering the promises of computational studies of sociological knowledge, and the contribution of this dissertation to the overarching sociology of sociology literature.
Questioning the Fragmentation of Sociological Knowledge

From the pioneering work *The Coming Crisis of Western Sociology*, written in 1970, to the posthumous book *Against Fragmentation*, published in 1985, Alvin Gouldner was on a theoretical quest to overcome fragmentation in sociology through Marxism. While the promises of Marxism were never realized, the issue of fragmentation addressed by Gouldner remains one of the most entrenched notions of what has been thought to be wrong with sociology and its knowledge production for half a century (e.g., Scott, 2005; Steinmetz and Chae, 2002; Walby, 2021). The paper by Holmwood (2010, p. 650) traces the idea of fragmentation even further, back to the notion of sociology’s disintegration in the 1940s, to the current version of self-subverting specialization. Thus, to investigate whether the corpora offer evidence that sociological knowledge is in a state of fragmentation was the main object of the first research question: *If fragmentation is understood as a historical process where the content of the knowledge produced in a discipline becomes increasingly disintegrated, do the corpora offer evidence that sociological knowledge is in a state of fragmentation?*

In general, the analyses indicate that the idea of acute fragmentation proposed by key figures in sociology – “that Sociology is losing its core, being hollowed out by fashionable themes” (cf. Walby, 2021, p. 323) – is not expressed in the corpora and that sociological knowledge does not seem to become increasingly disintegrated. However, in the literature review, it was argued that some studies seem to apply the notion of fragmentation as a synonym for specialization (e.g., Korom, 2020; Leahey and Moody, 2014), for which there is proof in these corpora.

In both Chapter 6, focusing on word constellations, and Chapter 8, dedicated to linguistic style, it was suggested that the analyses point in the opposite direction of fragmentation. For the former, this came in the form that several word constellations recurred over time, and in the latter, it was expressed in a set of persisting linguistic styles. However, the notion was most thoroughly investigated in the second analytical chapter, Chapter 7, where a structural topic model was generated on the abstract corpus covering the period 1980-2019. While a good share of these topics appeared to be more prevalent in either the 20th or the 21st century, the analysis suggested that a majority of the topics were rather well-distributed over time. The latter indicates that the dissertations in sociology defended in Sweden have tended to revolve around similar topics over time. This study thus sustains observations presented in previous computational studies of topics in sociology based on corpora of journal articles. In the paper by Moody and Light (2006, p. 68), the authors
suggest that sociological knowledge is not characterized by fragmentation but by “ebb and flow with few dramatic fissures”, with minor topical trends and a core of topics remaining over time. On a similar train of thought, Zougri (2019, p. 65) suggests that sociology is characterized by a “structural thematic isomorphism” based on separate national and international traditions sharing common topics with a “high level of consensus in the core”.

By suggesting the appearance of a topical core (supported by some private studies), the findings of this study differ from the only study of dissertations found in the literature. In their paper, Heiberger and colleagues (2021, p. 1167) selected the PhD level to specifically capture “the specialization strategies of young sociologists”, which are assumed to be “charged with reproducing or transforming the field”. The study presented in this dissertation and the study conducted by Heiberger et al. (2001) diverge on this point. Indeed, the analyses of abstracts for dissertations defended in Sweden, presented in Chapter 7, lean towards the former when suggesting that there is indeed a tendency for specialization over the 40 years analyzed, but this is coupled with a topical core. In comparison, the study of a selected corpus of dissertations defended in the US points to the latter by suggesting that no specific set of topics is promoted. Heiberger and colleagues (2021, p. 1167) further present the argument that sociological topics are specialized to an extent that “fuel the disciplinary lament of fragmentation”. However, there remains an unsettled debate amongst scholars in the field on whether specialization actually leads to fragmentation (cf. Scott, 2005) or if it is rather a necessity for scientific development (cf. Vanderstraeten, 2010).

Leaning on the work on the fractal theory of sociology introduced by Abbott (2001), Moody (2004, p. 215) argues that sociology has porous boundaries and that specialization does not necessarily generate division, but that the fast alteration of the content of knowledge “contributes to perceptions of fractionalization”. By inferring the results generated in the stylometric analysis conducted on the full-text corpus, this study suggests that underneath the level of topical specialization might lie more consistent forms and recurring representations of sociological knowledge. In the analysis presented in Chapter 8, it was suggested that the five linguistic styles can be seen as textual representations of different forms of sociological writing. While dissertations with either a distribution-oriented style or a concept-oriented style were used to juxtapose the most diverse dissertations in the full-text corpus, the majority of the dissertations analyzed fell within the three other linguistic styles – person-oriented style, relation-oriented style, and circumstance-oriented style – which appeared to be closer to each other and exemplify a type of ‘general sociology’ lying between the two aforementioned styles.
Therefore, the results presented in Chapter 7 and Chapter 8 suggest that even if there are, relatively speaking, extremes to be found within the topics and the linguistic styles modeled on the corpora, there seems to be in both cases an intangible ‘core’ that is more stable over time. One could further speculate on the relationship between linguistic styles and topics in the corpora of dissertations in sociology. Perhaps the former represents a limited selection of more robust and accepted forms of generating knowledge in sociology and the latter the current thematical trends, which are more fluent. This might then mean that a dissertation in sociology would have to limit itself to one of the linguistic styles that then, theoretically speaking, could be combined much more freely with a topic of choice fitting the interests of the study. However, at this stage, this is just speculation that would have to be investigated in depth with future inquiries.

Lastly, there is another pattern found in both Chapter 7 and Chapter 8, as well as a previous study, that can be related to fragmentation. In the structural topic model of abstracts based on dissertations addressing sociological terms, the analysis presented by Heiberger and colleagues (2021) indicates that the topic of ‘theory’ went from being the most popular topic in the early 1980s to continually declining as the 21st century unfolded. The authors suggest that “theory … has become an auxiliary tool for empirical graduate research and less of a thought style in its own right” (Heiberger et al., 2021, p. 1180). Similarly, in the abstract corpus analyzed in this dissertation, it seemed as if the dissertations over time appeared to abandon theoretical sociology in favor of empirical sociology. In the structural topic model, this was expressed in the steadily declining prevalence of the Theory/Knowledge topic from 1980 to 2019, and in the stylometric model as the complete disappearance of dissertations with a concept-oriented style in the 2010s. In the analysis of word embeddings, it was suggested that dissertations with a concept-oriented style were the ones primarily addressing the term sociology and words found related to it in the full-text corpus, such as science, theory, and philosophy.

Thus, the trend seems to be that theoretical studies, specifically meta-discussions of sociology, have become less of a focus for the dissertations over time. Within paradigm theory, one central characteristic of a pre-paradigmatic discipline that has not yet reached the stage of the Kuhnian ‘normal science’ is to have “continual philosophic discussion over fundamentals” (Kuhn, 1962, p. 159). In the former case, a scientific discipline “is barely distinguishable, if at all, from ‘the philosophy of’ the relevant subject” (Maner, 1970, p. 74). The abandonment of meta-discussion of sociology in the corpus could thus be interpreted as a step away from disintegration and fragmentation and towards integration and unification. This conjecture leads us to the next section, where a discussion of what the com-
putational text analyses of the corpora have to add to the paradigm problem in sociology.

Two Paradigmatic Images of Sociology

The sociology of sociology was established in the early 1970s by a set of works with titles referencing this notion, probably most clearly in Friedrichs’ *A Sociology of Sociology* (1970) or Reynolds and Reynolds’ *The Sociology of Sociology* (1970). The interpretation suggested in this dissertation is that the pioneering works, to a greater or lesser extent, had Kuhn’s paradigm theory as one of their main points of departure for conceptualizing the crisis of sociology of their time. The common thread here was that sociology was conceived to lack a unifying paradigm, which was coupled with an explanation of the discipline, part theoretical and part historical. Yet, the focus shifted from explaining the loss of a previous paradigm and replacement of a new single paradigm (cf. Gouldner, 1969), two competing modes of conducting sociology or dual-paradigm (cf. Friedrichs 1970), or a few co-existing paradigms or multi-paradigm (cf. Ritzer 1975). Based on the presented review of contributions to the sociology of sociology in the 21st century, it was further proposed that the paradigm problem still lingers in the literature. A recurring theme appeared to be that of “Sociology’s Big Divide” run “between a discipline that seeks to be a science and one that does not” (Turner, 2015, p. 289). This “methodological separatism” (Erola et al., 2014) has been given many names but is often referred to as “the methodological divide” (Schwemmer and Wieczorek, 2020) based on “the distinction of qualitative versus quantitative” (Abbott, 2001, p. 28). Thus, this methodological divide has been given extra attention in this dissertation, where it is interpreted as a version of conceptualizing sociology as a dual-paradigm science. The paradigm problem in sociology prompted a research question based on an operation of the theory able to trace potential textual representations of paradigms in the corpora (which was reiterated in the introductory section of this chapter): *Given that a paradigm denotes a crude replica constituted by reiterations of the same crude analogies in a discipline, do the corpora offer evidence that there are one, two, multiple, or no paradigms in sociology?*

Returning to Masterman’s (1970, 2009) reading of paradigm theory guiding this dissertation (see Chapter 3), a paradigm is said to exist if there is evidence of crude replicas of the paradigm in the form of crude analogies that are expressed in text. Following Abbott’s (2001) fractal theory of sociology, crude analogies are, in this study, mapped out through combinations of words in the corpus that in some way resonate with prevailing sociological dichotomies, nine of which are presented in Chapter 3. Since neither corpus
articulated the same combination of dichotomies, this study proposes that sociology does not appear to resonate with the image of a single-paradigm science (or ‘normal science’), which according to Kuhn (1962, p. 14) are disciplines based on “universally received paradigms”. This point is commonly reiterated within the sociology of sociology (e.g., Arluke, 2002; Brante, 1981; Eriksson, 1997; MacLean and Williams, 2008; Wadsworth, 2005), and has been supported in previous empirical studies (cf. Evans et al., 2016).

However, instead of presenting a narrative of complete fragmentation, the various analyses conducted – from Chapter 6 to Chapter 9 – all proposed that the corpora of dissertations in sociology in some way articulate one or several of the prevailing dichotomies. Thus, the notion of non-paradigm science does not appear to apply to sociology in this case. This position is further supported by inferring the observations made earlier that the corpora do not express a clearcut fragmentation (Chapter 6, Chapter 7, and Chapter 8), and that theoretical sociology has declined over the last four decades in the corpora in favor of empirical sociology (Chapter 7 and Chapter 8). These results go against some contributions to the literature reviewed, for instance, the paper by Moksony and colleagues (2014, p. 18) in which the assumption is that sociology lacks a paradigm, entailing that most “ground-laying work” of the discipline is not settled, which results in that “everything always needs to be stated anew”. Nevertheless, based on the interpretations of the corpora presented here, knowledge production in sociology appears to lie in between the two ‘extremes’ of non-paradigm science and single-paradigm science.

At this stage of scrutinizing the problem guiding this section, both the non-paradigm science and single-paradigm science have been rejected, and the issue has now become whether the corpora were found to resonate with a dual-paradigm science or a multiple-paradigm science. The answer provided to this question will have to be two-fold. This is because the analyses of the abstract corpus, Chapter 6 and Chapter 7, told a different story than those based on the full-text corpus, Chapter 8 and Chapter 9. Beginning with Chapter 6 and Chapter 7, the analyses conducted on the abstract corpus appear to sustain each other. In the first analytical chapter based on word constellations, Chapter 6, the clustered network of word correlations showed a separation between two communities that were interpreted to pick up words associated with the opposing poles of the three methodological dichotomies – science/humanities, explanatory/interpretative, and quantitative/qualitative – and of one ontological dichotomy – macro/micro. Added together, these were conceived to fit the notion of a division between a quantitative community (seeking to be a science) and a qualitative community (drawing on humanistic and interpretative traditions). In the second analytical chapter, Chapter 7, which was dedicated to delineating thematic arrangements, 18 of
the 20 topics generated by the structural topic model formed a network of two major communities, where the topics that were most strongly associated with the *quantitative community* were found in one community and those most clearly echoing the *qualitative community* in the other.

In other words, the analyses of both word correlation networks and the structural topic model (Chapter 6 and Chapter 7) appear to echo the idea of the methodological divide discussed in the theoretical literature (e.g., Abbott, 2001; Azarian, 2022; Erola, et al., 2014; Turner, 2015) and mapped out in the empirical sociology of sociology literature (e.g., Gartrell and Gartrell, 2002; Moksony et al., 2014; Schwemmer and Wieczorek, 2020; Traag and Franssen, 2016). Putting this result in the framework of paradigm theory, the analysis of the abstracts for dissertations in sociology generated a view of the discipline that is closest to that of a *dual-paradigm science* comprised of “two competing paradigms struggling for the mastery” (Masterman, 1970, pp. 73-74). When comparing the word constellations associated with each community, it was found that the word correlations adhering to the *quantitative community* were more prevalent, durable, and interconnected in the network. Thus, it appears that the *quantitative community* uses a more univocal set of concepts in comparison to the *qualitative community*. Further, the group of topics coupled with the *quantitative community* was more correlated. Thus, the cumulated results presented in both analytical chapters (Chapter 6 and Chapter 7) appeared to carry some evidence for the interpretation that the *quantitative community* expressed a more uniform set of crude analogies that could function as crude replicas for reproducing its paradigm than the *qualitative community* (Masterman, 1970, 2005). Perhaps a case can be made that this finding reflects the notion presented by Collyer (2013a, p. 349) that qualitative sociology tends to be based on small and loose “theoretical networks” and quantitative sociology on sturdy and intergenerational teams providing “substantial, on-going, in-house expertise”. Yet, this would call for more targeted analyses of metadata.

However, this rather neat image of sociological knowledge as the result of two opposing communities grounded on the master dichotomy of *quantitative/qualitative* is solely based on the empirical material of abstracts, which was suggested in the literature review to be more or less the standard practice of computational studies of sociological texts (e.g., Moody and Light, 2006; Seale, 2008; Zougris, 2019; Moody et al., 2022). When considering the results generated from the full-text corpus analyses, presented in Chapter 8 and Chapter 9, another narrative unfolds. Beginning with the stylometric analyses of Chapter 8, dissertations with either a *person-oriented style* or a *distribution-oriented style* were found to address words for *empirical inquiry*. Thus, they could be construed as representatives of the two sides of the methodological divide, yet when including the other three linguistic
styles a more multidimensional pattern emerged. For instance, dissertations with either a concept-oriented style or a relation-oriented style were construed to be theoretical rather than empirical. While the former was interpreted to express a focus on human agency and the micro level, the latter seemed to be associated with social structure and drawing toward the macro level. Thus, it was suggested that a more enhanced understanding of the separations between the linguistic styles in the full-text corpus could be found by inferring the epistemological dichotomy of empirical/theoretical and the ontological dichotomy of structure/agency to the divide of a quantitative community and a qualitative community (conceived to be constituted by the opposing poles in the dichotomies of quantitative/qualitative, science/humanities, explanatory/interpretative, and macro/micro).

Thus, the interpretation of sociology as divided by a methodological divide found in the abstract corpus was not as clear-cut in the full-text corpus. These analyses suggest that different types of crude analogies of sociology, operationalized in the form of sociological dichotomies, can be delineated in the corpora. Further, each of the five linguistic styles uncovered in Chapter 8 appears to reiterate a certain combination of dichotomies. Following Masterman (1970, 2009), this can be interpreted as an analogical pile-up, here in the form that the same dichotomies are repeated within the same linguistic style, which is the act of forming crude replicas of a paradigm. Thus, in this case, five co-existing paradigms appear to be articulated in the full-text corpus. The specific number, form, and content of these paradigms ought to be regarded as highly tentative at this stage, and would most probably vary in different corpora and contexts, which calls for further investigations.

Nevertheless, if looking solely at what the results of the computational text analyses based on the full-text corpus suggest, it would seem reasonable to describe the discipline of sociology as a multiple-paradigm science. In contrast to the “mature single paradigm science”, we would here find that “each sub-field as defined by its techni que is so obviously more trivial and narrow than the field as defined by intuition” and that “discussion on fundamentals remains” (Masterman 1970, pp. 74-75). Indeed, the most pervading finding of the analyses conducted on the full-text corpus presented here seemed not to be that there is a rather clear divide between the linguistic styles – whether it is between dissertations with a distribution-oriented style and a concept-oriented style or a distribution-oriented style and a person-

\footnote{Quite prophetic, the social sciences come up as an example of multi-paradigm science when Masterman (1970, p. 74) contrasts it with non-paradigm science: “This pre-scientific and philosophic state of affairs sharply contrasts, however, with multiple-paradigm science, with that state of affairs in which, far from there being no paradigm, there are on the contrary too many. (This is the present overall situation in the psychological, social, and information sciences.)"}
oriented style. The tendency found in Chapter 8 was rather that the linguistic styles articulated a more complex pattern, where two styles would be positioned on the same pole in one dichotomy and opposing poles in another. Thus, the linguistic styles appeared to transgress being sorted into a single ‘master dichotomy’. In other words, sociological knowledge, as it is presented in sociology dissertations in Sweden, does not seem to be a choice of two exclusive communities. Rather, there seems to be room for a plurality of paradigms – each with its more or less stable ontological, epistemological, and methodological traditions. Since the analyses of the full-text corpus suggested that there are a small and relatively fixed number of paradigms, it is proposed that this form of pluralism ought not to be conflated with relativism (cf. Lamont and Swidler, 2014).

The conclusion regarding paradigms is, thus, that the dissertations in sociology articulated a dual-paradigm science on the level of abstracts and a multi-paradigm science on the level of the full-texts corpus. It is here suggested that this discrepancy forms one of the principal roots for unravelling the overarching aim of this dissertation and, thus, calls for a separate section that allows for moving beyond the empirical study into a more speculative domain. Here, the results of the analytical chapter on word embeddings, Chapter 9, which has hitherto played an insignificant role, will be central to the explanation. However, before diving deeper into this puzzle, the next section will be dedicated to the last research question on the social conditioning of sociological knowledge.

The Local Conditioning of Sociological Knowledge

In a famous paragraph of Tristes Tropiques, Claude Lévi-Strauss (1955, p. 148) wrote that “One must be very naïve or dishonest to imagine that men choose their beliefs independently of their situation”. At least since 1929, with the publication of Karl Mannheim’s Ideologie und Utopie, this idea has formed a central tenet in the sociology of knowledge. Indeed, Mannheim (1960, p. 237) stated that the field “has set itself the task of solving the problem of the social conditioning of knowledge” and concerns itself with “the significance of the non-theoretical conditioning factors in knowledge”. The sociology of sociology studies reviewed was suggested to focus on two overarching forms of conditioning concerning sociological knowledge: history and geography.

The temporal dimension is prevalent in almost all empirical studies since the papers pay attention to how knowledge evolves. The spatial dimension, on the other hand, is more diverse – including global (e.g., Collyer, 2014;
Zougris, 2019), national/regional (e.g., Abend, 2006; Grothe-Hammer and Kohl, 2020), and local/place (e.g., Aaltojärvi et al., 2008; Bjarnason and Sigfusdottir, 2002) forms of conditioning. Based on the review, it was suggested that a fruitful contribution to the literature would be a study that can simultaneously map out how sociological knowledge changes at multiple localities within a broader national or regional setting. This is why this thread will be picked up in this dissertation focusing on departments of sociology located in Sweden to explore the social conditioning of sociological knowledge by studying how it is produced at multiple localities within a single nation over time. In the frame of dissertations in sociology, time is operationalized as the year a dissertation is defended, and locality as the place awarding the doctoral degree in sociology. Thus, the research question posed has been: Following the notion that the social conditioning of knowledge is expressed in the time and place of its production, do the corpora offer evidence that sociological knowledge is socially conditioned?

A tentative answer to the first part of this question, dealing with how the corpora are conditioned in terms of time, has been provided in the previous section on fragmentation but will now be scrutinized more thoroughly. Beginning with the methodological divide proposed to be one of the overarching patterns in the abstract corpus, it was suggested to be rather consistent over time, both in terms of how it was expressed in the clustered word constellations (Chapter 6) and in the clustered topic network (Chapter 7). Continuing with the structural topic model of the abstract corpus, the analyses of Chapter 7 suggested that a ‘core’ of topics consistently recurred over time, while the rest appeared to follow more defined historical trends since they were more prevalent during either the 20th century or the 21st century. Further, from the stylometric bootstrapped consensus network analyses of the full-text corpus presented in Chapter 8, it can be proposed that only dissertations with a concept-oriented style expressed a declining prevalence across the period, while the other four linguistic styles were fairly well distributed between 2000 and 2019.

Based on these results, it seems that trends play a smaller part than expected in the overarching story of how the content of the corpora is structured. However, one should be careful not to draw too strong implications from this tendency. Indeed, when looking underneath the aggregated level, more complex patterns tend to emerge. For instance, in Chapter 6, exploring the strongest word correlations within the abstracts for the dissertations (1980-2019), it appeared that the middle of the period (1996-2009) expressed the most unified language, with distinguishable themes like methods, theory, and data, as well as labor market, gender, and (welfare) policy. In contrast, the decades before were implied to have a more disintegrated language. This suggests that there are still historical variations in word usage and, thus, also
shifts in the way sociological knowledge is presented over time. Nevertheless, when considering the overarching patterns suggested to structure the content of the corpora, it seems as though time is less important than would have been expected from the perspective of more historically inclined studies of sociological knowledge (e.g., Crothers, 2008; Harley, 2008; Platt, 2008; Schrecker, 2008).

Besides time, it was proposed that many studies tend to stress the importance of spatial conditions of sociological knowledge, which makes up the other part of the research question guiding this section. Particularly, the focus is placed on measuring the effect of place on the corpora by comparing the universities producing the dissertations. This angle of the investigation was inspired by previous studies suggesting that there are “institutionally and geographically-tied” schools of sociology (Collyer, 2013a, p. 349) and that departments have “a distinct intellectual identity” that contributes to the overarching constitution of national and transnational communities of sociology alike (Bjarnason and Sigfusdottir, 2002, p. 255).

As was repeatably proposed in the analytical chapters, the methodological divide (dual-paradigm science), which is suggested to be one of the principal divisions in the abstract corpus (Chapter 6 and Chapter 7), appears to recur over time and be contingent on place. Indeed, when categorized by their university, two groups of dissertations were suggested to be more associated with the quantitative community – Stockholm University and Umeå University – and two with the qualitative community – Lund University and Uppsala University. Dissertations defended at the last university – Gothenburg University – were found in the middle of the divide regarding word correlations but more associated with the topics regarded as related to the qualitative community. That is, looking solely at the abstracts for dissertations in sociology, one could presumably predict that, for instance, when a dissertation in sociology is produced at Stockholm University and another at Uppsala University, the first would entail a more quantitative form of sociology and the other a qualitative form. Thus, in the case of the abstract corpus, it appears that this textual representation of sociological knowledge points in the direction that sociological knowledge is more conditioned on place than time.

When the relationship between place and linguistic styles was investigated in the full-text corpus, Chapter 8, a similar methodological difference between the universities was proposed. The two universities suggested to be more attuned to the quantitative community – Stockholm and Umeå – had a higher percentage of dissertations with a distribution-oriented style compared to those proposed to be attuned to the qualitative community – Lund and Uppsala – and the complete opposite pattern was found for the person-oriented
style. Yet, a more complex pattern emerged when adding the three remaining linguistic styles and inferring the sociological dichotomies they are suggested to articulate. Here, a plausible interpretation that was presented, but that ought to be read with great caution, is that dissertations from Stockholm and, to a lesser degree, Umeå are more inclined to produce sociological knowledge related to the quantitative, positivist, macro, and (together with Gothenburg this time) empirical poles. In contrast, dissertations from Uppsala and Lund seem to articulate sociological knowledge that is more qualitative, interpretivist, micro, and (mostly for Uppsala) theoretical in its content.

Thus, based on the analyses conducted on both corpora, it seems that local conditioning is a particularly strong indicator of what type of sociological knowledge is produced. This is expressed in the various forms of clustered networks of word correlations (Chapter 6), latent topics (Chapter 7), and linguistic styles (Chapter 8). However, this conclusion also ought to be drawn with caution since all results were not completely unanimous. For instance, in the analysis of word embeddings, Chapter 9, several measurements were conducted of the effect of place and linguistic style on key concepts in sociology (sociology, society, and a selection of sociological dichotomies). These suggested that the linguistic styles were better indicators than the universities for uncovering opposing positions in the corpus. Thus, it seems that the universities do not encapsulate meaning-making as such but rather the overarching approach to sociological research as expressed in the more abstract concepts of the methodological divide or the dual-paradigm science and multiple-paradigm science. When discussing this tendency in relation to the literature, there seems to be some confirmation that departments of sociology foster certain intellectual identities (Bjarnason and Sigfusdottir, 2002) or schools of sociology (Collyer, 2013a).

Based on the tendencies found when studying the local level, it seems necessary to address the national level since this is the most common aspect of spatial conditioning found in the literature. Indeed, the importance of the national context of sociological knowledge is a frequently revisited topic within the sociology of sociology (e.g., Erola et al., 2015; Heilbron, 2014; Platt, 2007b). For instance, the analyses of Mexican and American sociology journals performed by Abend (2006, p. 29) conclude that the countries represent two incommensurable “Styles of Sociological Thought”. A more moderate position is found in a study of American and British sociology journals (1974-2013) conducted by Zougris (2019, pp. 64, 82), where the author argues that the “fractal thematic landscape” of sociology is explained on the level of different national traditions, where some topics are nation-specific and others shared. Translated into the methodological framework of this dissertation, these studies pose the question of how strong the national con-
ditioning of sociological knowledge is when comparing the case of Swedish dissertations in sociology with previous studies of corpora in other contexts.

In a study of “all English-language research papers indexed by Sociological Abstracts from 1970 to 1990” conducted by Moody and Light (2006, p. 74, 82), it was presented that “a core of work on organizations, stratification, social institutions (family, science, religion), identity and difference (race, sexuality, language), and culture remains throughout the period under study”. This core seems close to what Crothers (2011, p. 5.5) has referred to as the “big categories”, namely “history and theory, economic and political, stratification, and family and cultural sociologies”. These themes are suggested by the author not only to be present in their study of British sociology journal articles (1966-2010) but also found in other empirical studies of international sociology outlets. For instance, similar themes can be found in another study of four mainstream British sociology journals (1999-2000) by Payne and colleagues (2004, p. 158), where an additional set of topics is also included that is constituted by “Education”, “Race and ethnicity”, “Gender”, “Criminology and law”, and “Research methods”.

Recalling the analysis of the topics modeled on the abstract corpus, given in Chapter 7, the core topics and the big categories can be found, although in slightly different combinations: Organization/Movement, Theory/Knowledge, History/Economy, Class/Inequality, Family/Background, and Space/Culture. Further, the expanded set of themes found by Payne et al. (2004) seems to resonate in the topics of Education/Field, Ethnic/Migration, and Gender, whereas a case can be made that research methods are included in Qualitative and Positivist/Risk. However, there is no equivalent topic for criminology and law, whereas Welfare/Countries seems to be a topic particular to the Swedish (and presumably Nordic) case. Further, the topic model of 60 topics of sociology-related theses defended in the US, 1980-2015, presented by Heiberger et al. (2021), includes equivalent versions of the 20 topics of the present study. In line with the results presented in this dissertation, the authors also suggest an increased prevalence of “topics related to the cultural turn”, particularly identity and qualitative methods, and “the decline in theory” (Heiberger et al., 2021, p. 1180). Worth mentioning is that among these 60 topics are a few that seem to be more specific to the US, such as “Race: African American” and “Race: Latinx” (Heiberger et al., 2021, pp. 1172–1174).

---

160 Perhaps a case could be made that the topic called Problems/Treat, which picks up words associated with social problems like violence, homicide, and alcohol, reflects the field of criminology.
Thus, there seem to be commonalities between the results of the analyses conducted on the abstract corpus in this dissertation and other studies. Worth noting is that these studies use the same type of empirical material, that is abstracts for publications in sociology around the turn of the 21st century, and similar types of methods, namely clustered word correlations and latent modeling. This suggests that in the case of sociology, studies of dissertations defended in Sweden (as well as the US) have proposed a similar thematic core to studies of international sociology journal articles. Yet, outside this core, there is some evidence for a few more nation-specific topics. This outcome echoes the tendency found by Zougri (2019). Worth noting is that the studies of comparison figuring in the previous paragraphs are all within the West. Therefore, in light of the national conditioning of knowledge, it can be suggested that, at least within the context of the West, national traditions appear to play a more limited role in the production of sociological knowledge than expected. This position is in favor of the institutional isomorphism thesis previously mentioned in this chapter, stating that national sociologies are increasingly becoming more alike (cf. Heilbron, 2014).

This section has been dedicated to discussing the last research question of the dissertations, dealing with to what extent sociological knowledge, as it is expressed in the corpora, can be said to be socially conditioned with a focus on time and place. The simple answer to the question is that the corpora appeared to rearticulate similar patterns of sociological knowledge over time and that these patterns appear to be structured by place, operationalized as the locality where the dissertations in sociology were produced. Indeed, these departments of sociology seem to uphold opposing poles of various dichotomies for delineating similarities and differences in sociological knowledge. Together these seem to form a totality on the national level.

Further, in a more speculative investigation, the results of this study were compared to the results of previous studies of other corpora produced in other contexts. Based on this comparison, it appears that the different national and international forms of sociology articulated similar types of topics and oppositions. This argument poses a question to the more historical strands of the sociology of sociology literature of whether the national conditioning of sociological knowledge production can be taken for granted. In addition, this potential ‘institutional isomorphism of sociology’ (cf. Heilbron, 2014) would speak for a pattern of unification in international sociology and against the crisis of sociology, in the version that ‘sociology is everything and therefore nothing’. This leads us to the next section, dealing explicitly with what the analyses performed suggest about the overarching aim of this dissertation.
Paradigm Pluralism – An Opening Out of the Crisis?

This dissertation has aimed to shed new light on the crisis of sociology by empirically scrutinizing prevailing disciplinary understandings of the state of its knowledge production through computational text analysis. Following the reasoning proposed in Chapter 1, one of the basic problems of the discipline leading to this aim was suggested to be answering the seemingly basic question, What is sociology? The most general of the computational text analyses performed in this dissertation seem to sustain the experience that sociology is not easily mapped out. For instance, if we return to the overarching word constellations found in the abstract corpus (Chapter 6), we are reminded that they were interpreted to address rather common words for humans, people, and general aspects of societies such as economy and politics, as well as standard words used in research to address methods. Thus, while there are examples of some concrete words that are recurring over time and place, the sum of these word constellations appeared to be both too diverse and too generic to easily summarize in a few sentences.

To expand on this issue, the word embeddings modeled on the full-text corpus, presented in Chapter 9, provided a unique opportunity to dive deep into the ways that the terms sociology, as in the name of the discipline, and society, representing its study object, were used in the actual text of the dissertations. Looking at the distributional meaning of the former, it was suggested that sociology is addressed with the broader domains of philosophy, science, and theory, which all appear to be located on the same abstract level. Based on this result, sociology could perhaps be explained as a discipline that is part science, part philosophy, which is dedicated to generating theory. However, while this might give a sense of sociology’s troubles, it is doubtful that this explanation is helpful from a pedagogical point of view. Turning to the term society, the analyses presented in the previous chapters showed that this term is often accompanied by words that relate to the economy, politics, culture, and societal comparisons. While the two former aspects of society are the research objects of two other social sciences (i.e., economics and political science), the latter two were interpreted as too generic to single out as something specific for sociology in comparison to, in the first case, the humanities as such, and, in the second case, the other sciences of society. All in all, by looking solely at the distributional meaning of sociology and society (how it is used in what is conceived to be textual representations of sociological knowledge), we have, in other words, not moved all that far from ‘square one’ as it was presented in the introduction.

However, when adding on the more theoretically-driven computational text analyses of the dissertations in sociology performed, it was sequentially re-
revealed that sociology appears not to be about ‘everything human’, as was queried in the introduction. Rather, sociological knowledge appears to be structured by recurring patterns of words that even extend the realm of a national context like Sweden, and empirical material like dissertations. To take one example, the analyses of word constellations suggested the same overarching pattern of similar combinations of words in this corpus – abstracts for dissertations in sociology defended in Sweden – as was found in previous studies generated from analyses of abstracts for international sociology journal articles (Schwemmer and Wieczorek, 2020). Similarly, as was discussed in depth in the previous section, the structural topic model presented in Chapter 7 of this dissertation was found to echo a similar ‘core’ of topics as in studies of American sociology-related theses (Heiberger et al., 2021) as well as international journal articles in sociology from the US and the UK (Crothers, 2011; Moody and Light, 2006; Payne et al., 2004). This recurring core was interpreted as evidence against fragmentation, which was proposed to be a prevailing understanding of sociology constituting the crisis narrative of sociology.

In other words, sociology appears to be ‘something’ rather than ‘everything’ when looking through the eyes of computational methods applied to texts produced within the discipline. This suggests that, at least in the case of Swedish dissertations in sociology, the most resigned commenters on the so-called crisis of sociology might be exaggerating the state of things. Moreover, the potential institutional isomorphism of sociology (cf. Heilbronn, 2014) could further be interpreted on the lines that the crisis might be an experience rather than an institutional reality, yet this would have to be investigated with other methods. Nevertheless, one problem remains to be sorted out, namely, how can an image of sociology as a dual-paradigm science be found in the abstract corpus whereas the image of full-text corpus seems to resonate with a multi-paradigm science?

Indeed, the idea that sociology is divided into two methodological camps – qualitative sociology and quantitative sociology – was found to be a prevailing understanding of sociology that is supported by empirical studies of abstracts (Schwemmer and Wieczorek, 2020), including the one presented in this dissertation. However, the analysis of word embeddings can add more nuances to why the case of a master dichotomy of quantitative/qualitative is not applicable to the full-text corpus. The results offered in Chapter 9 suggested that the dissertations were inclined to address different aspects of

---

161 An almost identical figure to the one presented in the chapter can be found in an analysis of “all articles published in journals classified as sociology in the Web of Science in 2010-2015” performed by Vincent Traag and Thoman Franssen at The Centre for Science and Technology Studies (CWTS), Leiden University, which is available online (https://www.cwts.nl/blog/article=n-q2v294 [accessed 2022-12-09]).
society which was patterned by linguistic style. The clearest contrasting point was found between dissertations with a distribution-oriented style, concentrating on societal comparisons and the economy, and dissertations with a concept-oriented style, discussing modernity and culture. However, the majority of the full-texts within the corpus were not found in these two linguistic styles but in the other three linguistic styles: the relation-oriented style, the person-oriented style, and the circumstance-oriented style. This suggests that most dissertations are actually found somewhere in the wide area of tension spanning between the two opposites that the two aforementioned styles can be said to represent.

This discrepancy between abstracts and full-texts poses a broader question to the field of whether or not abstracts are proper representations of the content of dissertations (and journals for that matter) or form a genre of their own. Perhaps it is the case that rather than articulating content that appears to be divided according to the dichotomies informing these analyses, the abstracts are to be conceived as presentations for groups of intended readers, which the authors sort according to the prevailing dichotomies (cf. Abbott, 2001). That is, while the abstracts can be easily divided by sociologists into either a quantitative community or a qualitative community, the same might not be the case for the actual dissertations. This could explain why the abstracts appear to follow a more univocal line of division than the actual full-texts of the dissertations. Put in the vocabulary of paradigm theory, while the dissertations in sociology would be presented along the lines of dual-paradigm science, they trespass the most simplified dichotomies in their written form. Again, the analyses of the latter suggest that most sociological knowledge writings are situated between the qualitative community and the qualitative community, and were interpreted to most closely resonate with multiple-paradigm science.

The coexistence of multiple paradigms would explain the experiences of crisis found in the current disciplinary debate of sociology (e.g., Ballantine et al., 2016; Holmwood, 2010; House, 2019; Lopereato 1999; Woolcock and Kim, 2000). To recall, on the practical side of the discipline, voices appear to articulate an intuitive idea of a science of society constituted by a seemingly universal sociological imagination (e.g., Haddad and Lieberman, 2002; Mollborn and Hoekstra, 2010; Misra, 2000). In contrast, the theoretical side of the discussions questions the fundamentals of sociology, which is captured in the idea of fragmentation or, in more positive terms, pluralism (cf. Scott 2005; Walby 2021). Further, the discrepancy between the tendency of dual-paradigm science on the level of abstracts and of multiple-paradigm science on the level of full-texts, suggests a blind spot in the disciplinary understanding of sociology and the prevailing dichotomies applied in this dissertation. That is, sociologists seem to imagine their world as dichoto-
mous and present their knowledge accordingly, presumably to become affiliated with one pole rather than the other, and catch the attention of particular groups. This is, of course, nothing unique for sociologists but might hold for all sciences or, if allowed to speculate even further, society in general.

It was suggested in the previous sections that the decline of theoretical sociology entails a move away from the image of sociology as a non-paradigm science based on ‘continual philosophic discussion over fundamentals’ (Kuhn, 1962, p. 159; Masterman, 1970, p. 74). To expand on this note, it is proposed that the governance of empirical sociology can be interpreted as a state where we have a more immediate sociological practice released from the necessity of having to theoretically define our discipline in each presentation and publication. When adding all threads lined out in this section into a single knot, then, it is suggested that pluralism is the most fitting concept for describing sociology as it is expressed through these computational text analyses of corpora based on dissertations in sociology (cf. Payne et al., 2004). Again, this form of pluralism should not be read as a synonym for relativism (in line with Lamont and Swidler, 2014). Indeed, in contrast to ‘everything goes’, the findings of the computational analyses presented here suggest that, while there appears to be a wide range of differences between the sociology dissertations in this data corpora, the results suggest the co-existence of multiple paradigms based on a tangible number of themes, linguistic styles, and aspects of society.

Based on the analyses conducted and the discussions presented, this dissertation responds to the prevailing understanding of a crisis in sociology by proposing that the discipline articulates, and ought to strive for, a form of paradigm pluralism. In this paradigm pluralism, ‘pluralism’ is encouraged but ought to be limited to only a few co-existing paradigms, the number and shape of which are flexible to the current state of society as a whole. Indeed, one problem with a discipline reflecting the image of a single-paradigm science, or even a dual-paradigm science where two antithetical paradigms compete for mastery, is that it has a low level of flexibility in questioning the fundamental image of the science. This is undoubtedly an ill fit for a discipline like sociology, which has expressed an inclination towards reflexivity (e.g., Bourdieu, 2001; Gouldner, 1969).

A risk of allowing only one fundamental image of sociology would be to reduce the width of potential analyses to a certain methodology, theory, or study object. As was proposed based on the full-text corpus analyses, the paradigms – which presumably take a slightly different shape in other corpora according to their context – carry divergent meanings of the fundamental image of sociology’s subject matter (cf. Ritzer, 1975). In response, the mode of paradigm pluralism here proposed allows sociology to keep its justifica-
tion as a “general social science“ that is “irremediably interstitial”, flexible, and inclusive by not excluding any major types of knowledge about society (Abbott, 2001, p. 6). Adding to the discussions of the social conditioning of knowledge, paradigm pluralism does not necessarily have to be realized within a single department of sociology or even a single nation (Abend, 2006; Zougris, 2019). Rather, following Abbott’s (2001) theory of self-similarity in sociology, by interpreting the results presented in this and previous studies they can be said to lean in a direction whereby a balance between the dichotomies upholding the paradigms might be upheld by multiple localities in one nation and multiple nations in the global sphere of sociology (see also Burawoy, 2016).

Contributions and Future Work

What are the prospects of the study forming this dissertation and what does it suggest about future studies in the sociology of sociology? In answering this question, the answer will naturally also touch upon implications for discussion in both the sociology of knowledge and computational social science. The section will begin with suggesting some of the main contributions of the study and end with drawing out potential future paths for the field.

Like the other contributions within this line of research, this dissertation has shown the potential of computational methods and digitized text to empirically scrutinize prevailing understandings of what sociology is in the discipline. As has been pointed out by sociologists beforehand, one of the main strengths of computational, in comparison to traditional, text analysis is its reproducibility (cf. Nelson, 2019, 2020). Further, techniques drawn from machine learning render us able to find unexpected patterns in the texts with less interference from the researcher. It is thereafter up to the researcher to evaluate how these patterns ought to be explained concerning the corpus subjected to analysis. Specifically, this study has shown the presumption that this type of method can be used for engaging in theoretical discussions (cf. Lindgren 2019, 2020). On this basis, it is argued that computational text analysis is too important a tool to dismiss for sociology of knowledge, and ought to be added to the general toolshed of sociological analysis along with the more traditional methods (e.g., interpretations of deep interview transcripts, statistical analyses of surveys, and narrative-centered historical analysis).

In contrast to the greater majority of the computational studies reviewed in this dissertation, focusing on either journal articles (e.g., Schwemmer and Wieczorek, 2020; Traag and Franssen, 2016; Zougris, 2019) or textbooks
(e.g., Crothers, 2008; Harley, 2008; Platt, 2008; Schrecker, 2008), the empirical material of the analyses performed here has been not only dissertation abstracts (see Heiberger et al., 2021) but also full-text versions of dissertations. This opens up the potentiality of mapping out a stream of sociological knowledge that can be conceived as partly educational and partly scientific, existing in a space between the suggested ideological tendencies found in the former, since textbooks are set out to hold the discipline together (Kuhn, 1962) and journal articles tend to be focused on a more specialist audience (Manterman, 1970).

As has already been stated, most computational studies of sociological knowledge are based on abstracts – which is also the tendency in broader and adjacent fields like scientometrics – presumably since they are easy and open to access through various digital archives, and come in a ready-made format for analysis. Thus, the comparison between abstracts and full-texts presented here is a new addition to the literature. Even if it was not anticipated, the consequences of including both corpora caused such tension that it forced itself into becoming one of the central themes of the study. This indicates that the problem of the abstract’s validity for representing the content of an article or book is heavily understudied and undertheorized in the field. The results of this study propose that the abstract should not be conceived as a ‘neutral description’ of a publication that can be mined unproblematically. It instead seemed as though the abstract functions as a sorting mechanism for positioning academic texts. In comparison to abstracts, full-texts were experienced to be much more challenging to work with but carry the potential of better exposing the arbitrariness and multiple position-takings (and contradictions) that a single text can contain.

Several reflective papers on computational text analysis and text mining within sociology have stressed “the purposeful weaving together of multiple methods” into a “methodological bricolage” (Bonikowski and Nelson, 2022, p. 1480). As was the case with including full-texts, the utilization of various computational tools proved to be important to get a more nuanced picture of the corpora. Thus, looking at the results generated in this dissertation, there seems to be value in applying multiple methods on datasets on different levels since they can expose the strengths and biases of each other. This will presumably lead to more comprehensible results from which interpretations and explanations can be formed. Indeed, if playing with the idea the dissertation had only relied on the word correlation networks modeled on the abstract corpus, it would have ended up with the conclusion that sociological knowledge is structured by a methodological divide (e.g., Schwemmer and Wieczorek, 2020). Similarly, had the same abstract corpus only been analyzed through the lens of a topic model, it would primarily have described the topics generated and their relation to a few metadata (e.g., Heiberger et
al., 2021; Zougiris, 2019). Worth noting is that this is a finding in itself, yet it poses a question reaching far beyond the scope of this dissertation: To what extent does the model generated by a specific computational technique unveil the ‘true structure’ of the data, and to what extent does it impose a ‘forced structure’ on the data?

Nevertheless, beyond providing thick descriptions of outputs based on the text models and interpretations of how well they align with claims made in previous studies, the most novel perspectives on sociological knowledge were generated by applying methods that were unvetted in the field, namely stylometry and word embeddings. Stylometry is frequently used in both digital humanities and computational linguistics (cf. Eder et al., 2016), but, to my knowledge, has yet to be introduced into the mainstream sociological literature. The linguistic styles deducted from the bootstrapped consensus network (Chapter 8) were found to be the most important assets for theorizing about paradigmatic representations of sociology in the corpus. One key benefit of this set of techniques is that they are based on raw word frequencies in a collection of texts, and need little to no intervention from the researcher. The question of whether or not the specific linguistic styles delineated in this study would show up in another study based on another empirical material is, of course, a completely open question. Nonetheless, that linguistic styles were found to be an effective way to move beyond deduction and map sociological knowledge, so to speak, in situ, is one of the findings that this dissertation proposes as a viable way forward for not only the sociology of sociology but the sociology of knowledge as a whole.

Word embeddings, on the other hand, have started to find their way from computer science research into sociology journals to study various cultural phenomena (e.g., Kozlowski et al., 2019; Stoltz and Taylor, 2021), yet, again to my knowledge, this technique has not been used for studies of sociological knowledge (see Chapter 9). An intriguing aspect of this method is that the embeddings are modeled on the local context of the words rather than the global word frequencies – the hybrid GloVe model is an exception (cf. Pennington et al., 2014). This characteristic opens up a new opportunity to find and interpret distributional meanings of words based on how they are used in the corpus rather than assuming their meanings. Word embeddings generated on the full-text corpus with the word2vec skip-gram model (cf. Mikolov et al., 2013) were able to provide a closer look at the usage of key concepts in an area of interest. In this study, they promoted, for instance, the interpretation that ‘sociology’ is a term primarily embedded in theoretical studies (in contrast to empirical studies), and what overarching societal dimensions are evoked when the term ‘society’ is used. Following more recent developments in word embeddings, the next natural step of development would be a transition to transformer models, such as BERT, due to their performance in
various benchmark tests (cf. Devlin et al., 2018), a transition that has already begun in sociology (e.g., McLevey et al., 2022).

Building on the insights generated in this dissertation, there are three paths set out for future studies of sociological knowledge. While the paths are in no way unique to the literature, the overarching analytical framework utilized in this dissertation suggests new and sociologically imaginative ways of how to address them that can be expanded further. First, this dissertation makes a genuine attempt at combining computational methods with the interpretive perspective for the sake of generating theoretically informed descriptions of sociological knowledge. Sociological knowledge was primarily conceptualized through paradigm theory and supplemented by prevailing dichotomies in the discipline, yet there are several dimensions of sociological knowledge worth investigating further. For instance, it is possible to build on Philipp Korom’s (2020) study of the most cited sociologists within the discipline and how these changed in the 1970s and 2010s. One example would be to perform computational text analyses on the works considered to be most influential by sociologists and investigate how assumptions in what counts as ‘good sociology’ change over time. Another related angle is to build on reception studies within sociology (cf. Baehr, 2016), to investigate the reception of sociological theories by comparing the content of the text presenting the theory and texts applying the theory. This train of thought could also be applied to the sociology of translation (cf. Heilbron, 1999) to model and compare original texts and translated texts. In both the latter cases, this would involve asking questions like what parts of the text are emphasized, and what parts are ‘forgotten’ (e.g., Ginnerskov, 2021).

Second, the particular case of dissertations in sociology can be investigated further in at least two ways. One is to add more disciplines to investigate how sociology dissertations look in comparison to, for instance, dissertations in economics and political science or representatives from the humanities and from the natural sciences, respectively. An intriguing work to build further on with more content-based computational analyses is found in a study by James Moody and Ryan Light (2006). In their study, the authors suggest that sociology is less bounded than its neighboring discipline and has moved away from topics related to social processes towards those related to social problems. Another way to deepen the study of sociology dissertations would be to consider countries other than Sweden. There are, for instance, studies of sociology in the Nordic countries that come with certain ideas of what the sociological knowledge of this region is about that can be scrutinized (e.g., Allardt, 1993), and text analyses based on hand-coding already lay a good foundation (e.g., Erola et al., 2014). Further, there is work done stating the importance of comparisons between sociological knowledge production in the global north and the global south (cf. Collyer, 2014). However, these are
just a few examples of all imaginable combinations of countries within global sociology.

Third, while abstracts were shown to be fruitful empirical material for delineating word constellations (Chapter 6) and thematic arrangements (Chapter 7), this dissertation has argued for the value of conducting computational text analyses of full-texts. However, the analyses were not set out to perform systematic benchmark tests to evaluate to what extent each corpus could perform different tasks, since this was not the aim of the dissertation. However, this allusion is supported in studies based on other scientific outlets. To take the probably most convincing example, a study in computational biology compared 15,000,000 full-texts and abstracts for the same journal articles and showed that full-texts outperformed abstracts in all benchmarked cases (Westergaard et al., 2018). Thus, the way forward in terms of data appears to be the whole full-texts like the topic modeling conducted on “all JSTOR sociology articles from 1890 to 2014 (142,040 articles, 157 journals)” presented in a working paper by Adel Daoud and Sebastian Kohl (2016, p. iii).

However, as was shown in the literature review, there is a risk that journal articles could be deemed to be the most dominant in the field, which comes with the risk of giving a one-sided view of the discipline. Further, the more steps one takes outside the, say, 20 core sociology journals, the fuzzier the disciplinary borders become, making it questionable what, for instance, the 157 journals represent. While dissertations make up an obvious candidate, we can, in terms of education, also think of theses on lower levels, like Master’s theses and Bachelor’s theses, or why not more systematic computational text analysis of textbooks or course syllabuses? On the research side, we can, for instance, think of monographs, anthologies, research applications, and research reports.

Fourth, this dissertation showed that a combination of methods gives a more multifaceted view of the corpus being analyzed, where certain interpretations of what is seen can be both strengthened and refuted. While topic modeling is becoming a formalized framework that has begun to find its way to the core sociology journals, where even an impressive study of 80,000 sociology-related dissertations has recently been conducted (Heiberger et al., 2021), the most interesting results in this dissertation were generated by adding stylometry and word embeddings. These two types of methods are not used in sociology yet and, thus, require more creativity and patience, even if they open up for a more interpretative take on computational text analysis, as this dissertation has hopefully shown. This methodology is not at all limited to studies of sociological knowledge but can, for instance, be built upon to study other cultural phenomena. A good example that resonates with this dissertation regarding how word embeddings can be utilized is a study by Austin C. Kozlowski, Matt Taddy, and James A. Evans (2019) that analyzes
how markers of class shift amidst economic transformations through word embeddings modeled on millions of books published over 100 years. Moreover, there are also a wide number of other techniques not applied in this dissertation to be found in machine learning and computational linguistics, the potential of which is unknown to me, which are worth exploring.
Acknowledgements

This dissertation is the result of a set of uncoordinated events that are hard to describe without narrating a story. I believe someone once said there are two kinds of stories: a hero goes on a journey, and a stranger comes to town. While both illustrate a disruption of normality, rupture appears to hit from opposite angles. In the first story, the relatable is affected by the unknown; in the second, the unknown affects the relatable. Within the frame of the dissertation project, I have read hundreds of acknowledgment passages written by previous PhD students in sociology defending dissertations between the 1950s and the 2010s. At least for this corpus, only one of the two stories is called upon to describe the PhD process, namely the journey. Yet, in contrast to the traditional hero, we find an indebted and malleable antihero navigating the wine-dark sea of academia with the hope of a liberating degree on the horizon. Perhaps a sociological explanation for why the story of the stranger is absent can be found in the low status of PhD students at the university: we have neither the power nor the influence needed to disrupt the town we entered.

The journey of this dissertation began when I pitched an idea for my coming bachelor thesis to Donald Broady after one of his lectures. This encounter led to me meeting Mikael Palme who encouraged me to pursue a master’s in sociology of education and hired me as a research assistant to conduct empirical research for his project on the internationalization of education in Sweden. This turn of events ignited my sociological imagination by putting the theories of Pierre Bourdieu into practice. The job also gave me access to an open office space where my desk was seated next to a visiting PhD student from Denmark whom I would happen to wed a few years later. Thus, I want to thank all my colleagues and friends working within the Sociology of Education and Culture (SEC) at Uppsala University, without whom I would not have pursued a PhD in sociology.

The real journey was initiated during the application interview at the Department of Sociology. I was welcomed by Hannah Bradby and Fredrik Palm, who since that day have been a source of support on various levels. My deepest gratitude, however, goes to my main supervisor Sandra Torres for pouring bucket after bucket of blood, sweat, and tears – I am the sole
source of the latter – into this project. Sandra, a rare quality of yours is that you never uttered your true opinion on what topic or focus the dissertation ought to have. Rather, you allowed me to embark on my own journey and explore multiple angles even though time was ticking faster by the minute. I would be a liar if I said that I left every supervision session feeling completely empowered, yet your skills in all aspects of academic writing and beyond have truly shaped my professional self for all futures to come. While Sandra played the strict father in this saga, teaching his child how to become independent with tough love, my second supervisor, Johan Wickström, took on the role of the caring mother. With optimism, empathy, and both feet burrowed deep into the ground, you made sure that I took care of myself and steered my focus back to the bigger picture. I treasure our times in the old departmental archives below the Karin Boye Library and how your expertise in history and education came to shape the dissertation.

Albeit only participating in the first phase of the journey, my initial main supervisor, Patrik Aspers, also left a clear mark on the dissertation project. This can be summarized in a line he uttered that would become formative for how the project could turn from ethnographic fieldwork to machine learning: “You cannot generate new insights into sociology by applying the tools of sociology.” During these initial years, the theoretical focus of the dissertation would form from a seed planted by Vessela Misheva in her philosophically inclined courses on the theory of science and classical sociology. Further, from 2016 until today I have cherished all colleagues who contributed to insightful discussions at the Cultural Matters Group meetings over the years. A special shout-out to Tora Holmgren as well as to David Redman, Magdalena Kania Lundholm, Dominik Döllinger, Sebastian Abrahamsson, and Mara Törnqvist for dedicating your time and work to CMG.

I also want to emphasize the companionship offered by my fellow doctoral candidates on our parallel journeys. This tribute is extended to a large group of great people centered around the PhD cohorts of 2015 and 2017 at the Department of Sociology and the diverse lot connected to the Uppsala Research School in Subject Education (UpRiSE) spread out across the university. At too many occasions to mention, you managed to turn working duties into joyous events. I am particularly thankful to Miia Bask and Sonia Kollner for acting as opponents at my mid-term seminar and for all the valuable comments you provided. The same kind of gratitude goes to the opponents of my final seminar, Simon Lindgren and Hannes Landén, who did a tremendous job in pointing out issues in the manuscript, which I, hopefully, managed to resolve for this final version. A special thank you is sent to Ilkka Mäkinen for carefully reading through the full manuscript (twice!) and providing thorough comments which truly enhanced the dissertation in several ways. This was not an effortless task, which can be sensed in the ambig-
uous words he uttered after delivering his sound critique: “This is a piece of work! I feel like I should receive a medal or something for reading it...”

The methodological shifts of the project have affected how the path of the journey, inside as well as outside the borders of Sweden. First, I want to thank all sociologists at an anonymous department that agreed to be interviewed and opened up their seminar rooms and classrooms for me to conduct participant observations. Second, I want to express my gratitude to the people helping out on the historical side of the project. RC08 History of Sociology at the International Sociological Association (ISA) has provided a great forum for discussing all the nitty gritty details in the world of sociology. Beyond this group, I want to thank you Hedvig Ekerwald, Gunhild Hammaström, Britt-Marie Johansson, Anna Larsson, and Per Wisselgren for all the insightful discussions on sociology in Sweden. I also want to thank Helena Olsson for opening the gates to the archives in Uppsala and Rasmus Axxér for aiding in digitizing the dissertations. Third, for the technical side of this project, the international Text Analysis Research Group, housed at the Institute of Analytical Sociology, Linköping University, has been a valuable source of inspiration. In playing the important role of the co-brainstormer, Roland Adorjáni, currently at the University of Amsterdam, helped steer the computational part of the project on its right course.

A journey cannot be accomplished without material means. I want to express my gratitude to Norrlands nation and Seth M. Kempe Minnes stipendiefond, together with Norrlands nations jubileumsfond (B), for providing a generous source of financial support throughout this journey. In the later stages of finishing the dissertation, aid to which I am deeply grateful was sent by Uddeholms forskarstipendium, Värmlands nation, and Carl Gustaf och Carl Cervins stipendiefond, Smålands nation. I also want to thank CMG and the Swedish Sociological Association for playing a substantial role in realizing my attendance at the XX ISA World Congress of Sociology in Melbourne, Australia. Further, two valuable one-week writing retreats were provided by Bergmansgårdarna, at Ingvar Bergman’s Dämbo on Fårö, and Harald och Louise Ekmans Forskningsstiftelse, at Sigtunastiftelsen Hotell & Konferens in Sigtuna.

Outside of academia, my family from the far north of Sweden has cheered from the sideline. I want to thank my dear parents, Fredrik and Annacarin, for doing their best to help out in any way possible and for being completely indifferent to my choice of work and studies yet supportive of whatever I do (or, in the case of my dad, after I quit doing it, that is). I also want to thank my siblings Johan, Simon, Julia, and Emanuel and their families as well as all my friends for serving as a stable foundation throughout the ups and downs of this journey. A shout-out to Johan and Helena in Stockholm and
Janne and Peter in Copenhagen for always keeping your homes open for spontaneous visits. In the early days of the project, when learning the craft of programming, Emanuel helped out with Python code and, together with Johan, aided me along the way to self-sufficiency (i.e., individual dependence on stack overflow). A special token of gratitude is sent to my incredibly generous parents-in-law, Bente and Jesper, for repeatably traveling to Uppsala on short notice to aid in times of trouble and for providing a haven in Kolding, the Tuscany of Denmark.

Finally, I can hardly put into words how dull and desolate this journey would have been without the support of my closest confederate, Mette. Like a rightful stranger, you entered my small town and shook the normaley out of its core. You are a true trailblazer, a warmhearted force of nature leaving loving marks wherever you travel. I am proud to call you my wife and so grateful for you and the beautiful boy you gave us. Aske, you are the apple of my eye – a bright child with a heart of gold. My proudest achievement in this journey is not some imaginary dragon I managed to slay while balancing on a web of deadlines. It is to have put you first at all times – however rough the circumstances were. I dedicate this book to you and your mother: the perfect antidote against succumbing to the world of letters.

*Josef Ginnerskov*

Uppsala, January 2024
References


https://doi.org/10.1088/1742-5468/2008/10/P10008

300


304


Lindgren, S., 2019. Hacking Social Science for the Age of Datafication. jdsr 1, 1–9. https://doi.org/10.33621/jdsr.v1i1.6


308


Appendix A: Dissertations in Sociology Defended in Sweden

This appendix aims to provide a context to the case of study, namely dissertations in sociology defended within the borders of Sweden, meaning at a Swedish university or university college.

Most case studies of sociological knowledge within a specific locality target places that in present historiography are seen as particularly significant. They can, for instance, be groundbreaking pioneers like the Chicago school (e.g., Abbott, 1999; Owens, 2012) or influential in contemporary developments like the cultural turn and the Birmingham school (Webster, 2004), or associated with a historically recognized sociologist like the Leicester Department with Norbert Elias (Goodwin and Hughes, 2011). In contrast, there are few sociological works grounded in empirical investigations of marginal and seemingly irrelevant localities in the world of sociological knowledge. The focus of this dissertation is a seemingly small and internationally insignificant country in the world of sociology. By focusing on the fringe of global sociology, this design aspires to make a move away from the major strands of the history of sociology by generating complementary maps of how sociological knowledge becomes actualized.

This dissertation is based on dissertations in sociology defended in Sweden. Following the literature reviewed, the context of a country matters for the production of sociological knowledge (cf. Abend, 2006; Bjarnason and Sigfusdottir, 2002). Sweden is undoubtedly a small and more peripheral country in the global discipline of sociology when compared to countries like the US or the UK (cf. Larsson and Magdalenic, 2015), which have often been subject to analysis (e.g., Gartrell and Gartrell, 2002; Grothe-Hammer and Kohl, 2020). As discussed in the methodology chapter, the fact that the dissertations are written in both Swedish and English also entails some limitations, meaning that compromises had to be made. Thus, from the get-go, it is recognized that these corpora might not be the typical case for studying patterns and trends in international sociology and the results should accordingly be read with this in mind. That being said, Sweden was selected due to its semi-peripheral position within the global system framework of sociology since it could add a divergent case to the general tendency, which has been
to prioritize Anglo-Saxon journal articles found in the literature. At the same time, research suggests that international sociology is subjected to an institutional isomorphism (cf. Heilbron, 2014; see also Zougris, 2019), making it more alike worldwide (see also Sztompka, 2009). This point will be discussed more thoroughly in the section dealing with the social conditioning of knowledge, where the similarity between the Swedish dissertations and international journal articles is addressed and discussed.

Sweden is a good example of a small country without a clear sociological tradition but with a concrete genesis. The authors of the most recent historical study of Swedish sociology, Larsson and Magdalenić (2015) propose that there never existed a distinct “Swedish sociology” since Swedish sociologists predominantly imported and modified foreign sociologies. In contrast, sociology in Sweden is defined by its early connection to and gradual disconnection from state policy and its unsettled history between American and European sociology (Allardt, 1993). Many Swedish sociologists and historians (Eriksson, 1994; Larsson, 2001; Wisselgren, 1997) describe a paradigmatic start in the form of an American neo-positivism in the middle of the 20th century that successfully erased all preceding European sociologies for about 20 years to come. This includes a Swedish sociology developed half a century earlier by Gustaf F. Steffen, the first Professor in Sociology and Economics, as well as the sociological knowledge developed for decades under philosophical departments at Swedish universities (Ginnerskov, 2021; Larsson and Magdalenić, 2015; Wisselgren, 1997).

Following the “student boom” of the late 1960s and onward through a crisis as well as a revitalization, sociology in Sweden is conceptualized to have gradually become increasingly European (i.e., “returned” to some parts of its “neglected past”) and plural in regards to epistemology (Fridjönsdóttir, 1987). Today, sociology and sociologists can be found at thirteen Swedish universities. While only five of these universities have a department of sociology, ten can publish sociology dissertations under the name of their university.

As stated in the footnotes in the analytical chapters that comprise the bulk of the empirical part of this dissertation, this study is neither about “Swedish sociology” nor PhD programs in Sweden. In contrast, the dissertations analyzed in the previous chapters are approached as “products” of sociological knowledge in their own right. Thus, the Swedish case has not been at the forefront of this dissertation even though this is a dissertation in sociology that will be defended at a Swedish university. This is why this dissertation – as a product of sociological knowledge in its own right – has relegated contextually important information about Swedish PhD programs, and the university system that is the backdrop for the dissertations that have been the
data upon which the computational analyses deployed earlier have been based, to this Appendix.

Thus, the purpose of this appendix is to provide the national context surrounding the dissertations in sociology by offering the kind of nationally relevant backdrop information needed to understand the Swedish PhD system. One of the most important contextual pieces of information is that there are nationwide legislations in Sweden on research education that control the production of dissertations. These determine which universities are allowed to award the *doctoral degree* and the *licentiate degree* and in what *research subject*. Two major reforms must be mentioned, the first in the 1870s and the second in the 1970s since they served to formalize what a dissertation is in a Swedish context. As for the case of sociology, all sociology dissertations within the period 1949-1999 were defended at one of five universities, each with a department of sociology, which were the only institutions awarding *licentiate* and *doctoral degrees* in sociology. At the dawn of the 21st century, five newer universities (none with a department of sociology) were also granted the right to give the two research education degrees in sociology. At this time, research subjects closely related to sociology also made an appearance.

In this respect, it is important to note that all sociology dissertations defended during the period 1949-1999 in Sweden are dissertations defended at a time when the five departments of sociology stood for all publishing of sociology dissertations. These are the five departments that have been operationalized as ‘place’ in the analytical chapters (i.e., in chronological order, Uppsala, Lund, Stockholm, Gothenburg, and Umeå). There is, however, more to Swedish sociology than these ‘spatial configurations’ which is why this Appendix includes an abridged presentation of the history of this discipline in this part of the world.

In Sweden, the state grants the right for a higher education institution to award academic degrees (*akademiska examina*) in an academic subject (*universitetsämne*), which are given at three different levels: *bachelor’s level* (*grundnivå*), *master’s level* (*avancerad nivå*) and *doctoral level* (*forsknivå*) (Utbildningsdepartementet, 2003, p. 38ff). For a manuscript to be published as a dissertation it has to be defended *viva voce* in a public setting within the framework of education at the doctoral level leading to one of the two diplomas: *licentiate degree* (*licentiatexamen*) or *doctoral degree* (*doktorsexamen*). Far from all disciplines and fields of study are represented as academic subjects and the ones that are most often only exist at the bachelor’s and master’s levels but not on the doctoral level. In the latter case, a subject is referred to as a *research subject* (*forskarämne*). Thus, a sociology dissertation is in this study conceived as the book defended at the end of a
research education leading to a licentiate or doctoral degree in a research subject including the term ‘sociology’.

During the 20th century, Swedish research education went through a set of changes that altered the context of producing sociology dissertations on more than one occasion. Since these changes will have to be taken into consideration in the analyses, an initial schema of the Swedish research education becomes inevitable. “The dissertation” as we know it as a social entity in Sweden today, has roughly 50-150 years on its back depending on the degree of formalization one refers to (e.g., Utbildningsdepartementet, 2003, p. 38ff). In previous centuries, students who had received a bachelor’s degree (baccalaureus 1447-1853, or kandidatexamen 1853-1870), could continue to study four years to receive the highest degree (at that time), master’s degree (magisterexamen). The dissertation emerged with the first step towards a formalization of Swedish research education, which can be traced to the realization of the so-called examination statue in 1870. This year, the master’s degree was replaced by the doctoral grade (doktorsgraden), and the licentiate degree (licentitatexamen) was added in between the bachelor’s degree and the doctoral grade as a preparatory step to the latter. To situate this date historically, women were not allowed to take part in university studies until 1873 (even if a few women from 1871 onward trespassed this restriction by receiving a letter of exception from the King of Sweden).

It would take almost a century before the second formalization of research education took place (e.g., Andrén, 2013, p. 63ff). In a Commission of the Swedish Research Education presented in 1968 called (Palmes) Universitetskanslerämbetets arbetsgrupp för fasta studiegängar (UKAS/PUKAS), which was approved by the Swedish parliament in 1969, the requirements for licentiate and doctoral dissertations changed immensely. In this Commission, it was concluded that the Swedish licentiate dissertation (from 1870 onward) was correspondent to the American Ph.D. dissertation and the Swedish doctoral dissertation demanded approximately twice the work. Starting in the early and finishing in the late 1970s, the doctoral grade was successively replaced by the doctoral degree (doktorsexamen) and with it a vastly reduced doctoral dissertation, now equivalent to the Ph.D. dissertation. At this time, licentiate degrees became almost obsolete, only to return again in a condensed shape in the 1980s, representing half of the doctoral degree and half of the new doctoral dissertation (i.e., half a Ph.D.).

For conducting a study of sociological knowledge, it is important not only to depict its present state but also to trace its multiple histories. As for Sweden, five sociology departments located at the five biggest and oldest universities in the country were for half a century, 1949-1999 (see Table 1), the sole producers of dissertations in sociology. Starting with the first sociology dis-
ertation defended in 1949, around 800 dissertations in total were published during these fifty years. Two years prior, in 1947, the first two departments of sociology were established and within one and a half decades all five universities in Sweden at that time had institutionalized a sociology department. Not only did each university have its own department – four out of five also had an affiliated university college in the 1960s. In most cases, these university colleges hosted, financed, and educated licentiate and doctoral candidates. Yet, when it came to the dissertation – a requirement for being awarded a doctoral and licentiate degree – it had to be defended at one of the five main universities. Thus, licentiates and doctorates could have pursued all their years of education at a university college but ended up in the dissertation series of another university following a two-hour public defense.

<table>
<thead>
<tr>
<th>University</th>
<th>Department of sociology</th>
<th>Sociology dissertations</th>
<th>University affiliation</th>
<th>University college</th>
<th>University</th>
<th>Sociology dissertations</th>
</tr>
</thead>
</table>

Table 1. A timeline of universities in Sweden granted the right to produce dissertations in sociology.

In Table 1, a timeline of the establishment and relations between universities and the affiliated university colleges, as well as the year the universities’ first dissertation in sociology (either doctoral or licentiate) was defended, is shown. These relations changed in the 1970s when the university college in Linköping became an autonomous university of its own and the other three affiliated university colleges became independent university colleges. Yet, the latter still needed their old host university for producing dissertations until 1999 when they were granted university status and, thus, gained autonomy. The 1990s also entailed a shift in the distribution of state-funded research resources where the grip of the old universities loosened and funds were spread out to all HE institutions (e.g., Andrén, 2013, p. 53ff). This change was especially favorable for the biggest and most established of the new universities: *Karlstad University, Linnaeus University, and Örebro Uni-*
versity. At this stage in time, “the big and old five” consequentially lost its monopoly for producing sociology dissertations.

If all sociology dissertations were defended at a sociology department in the research subject sociology during the 20th century, we would have ultimately would have dealt with approximately 1200 units. While the optimal cartography for this case should have mapped all dissertations within the social sciences defended in Sweden throughout the 20th century – at least 10,000 unities (a few hundred pages each) dispersed over 100 years – such a study design is, unfortunately, not possible to exercise in practice. To start with, the total corpus is huge and the information is neither digitalized nor ordered. Instead, there are several complications in just finding what books count as dissertations and demarcating what research subject they belong to. Further, there is an ever-greater loss of information in the early decades of the period, where uncertainty remains if you can find a copy of each book or if you have to rely on simple variables found in university archives such as author, title, and a one-word library code depicting the content of the book.
Appendix B: The Two Corpora and Their Limitations

The goal of this appendix is to discuss the character, coverage, and limitations of the two corpora. Following the bibliography of dissertations in sociology in Sweden created within the frame of this project, 1004 dissertations have been defended over 67 years (1951-2019). Looking at the periods in focus for this study, there are a total of 871 dissertations defended between 1980 and 2019, and 519 dissertations defended between 2000 and 2019. The abstract corpus covers 815 dissertations of the first periodization (i.e., roughly 94 percent coverage), and the full-text corpus 380 dissertations of the second periodization (i.e., around 73 percent coverage). The two corpora share the characteristic of having better coverage in the later parts of their respective time periods.

Figure 32. Time series of the stock of dissertations in sociology sorted by physical books, abstracts, and full-texts. Absolute numbers.
In Figure 32, the annual production of dissertations in sociology is presented according to the available data types: physical books, abstracts, and full-texts. Following the line of the physical books, we can gather that the 1950s and the 1960s constitute a time characterized by a small and stable production. From the year 1970 until around 1996, we see a shift where the number of dissertations produced each year increased by a factor of ten, only to almost two-fold again in 1996 until around 2013. Thereafter, we see a stable decreasing trend that by 2018-2019 has returned to the values of the 1980s. This trend suggests that the sociology dissertation as a genre in its own right became established in Sweden around the second half of the 20\textsuperscript{th} century and presumably had its heyday around the turn of the 21\textsuperscript{st} century.

When focusing on the line of the abstracts, we find almost complete coverage from the mid-1990s to the last year in the sample, 2019. In the 1980s and the early 1990s, on the other hand, the abstract corpus is less far-reaching. This is especially the case for the years 1988-1995, where the average number of dissertations being represented by an abstract is about three-quarters (76 \%). The simple reason is that these dissertations were coupled with abstracts, neither in digital archives nor inside of (or attached to) the physical books. One leading assumption for why there is worse representation in the 20\textsuperscript{th} century compared to the 21\textsuperscript{st} century is that different universities had diverse traditions for what metadata a doctoral dissertation ought to include and that this has become increasingly standardized over time.\textsuperscript{162} Following the trend in the number of dissertations in sociology defended in Sweden, it is also worth noting that the abstract corpus is skewed in favor of the period 1995-2015.

Turning to the line of the full-texts in Figure 32, we see that this line seems less well represented. Overall, there is better coverage in the 2010s than the 2000s, where the best period is found around the year 2012. Thus, the most obvious limitation of the full-text corpus is the beginning of the period, when there ought to be the largest production of dissertations but we find the lowest number of full-text per year. While over 90 percent of the dissertations defended during the 2010s are in the corpus, only roughly 50 percent of the dissertations defended in the 2000s were coupled with a full-text copy. With this follows that the full-text corpus is tilted towards the middle of the period (2008-2012) when it actually ought to be skewed to the start as this was

\textsuperscript{162} A multitude of propositions and reforms affecting PhD education in a more nationally standardized direction was suggested and implemented by the Swedish state during the period 1990-2019. At the same time, the number of universities in Sweden doubled from 1989 to 2017. (See https://www.vr.se/download/18.4dd26b09169ebe0ddda652/1555327344453/Reformer-inomforskning-och-forskarutbildning-1990-2017_VR_2018.pdf [accessed 2023-02-23])
when the highest annual number of dissertations were produced (2000-2004).

While it is hard to phantom what consequences the uncovered biases in the corpora will have for the analyses, it ought to be kept in mind that it is possible that the full diversity of each corpus is not captured since the early period is not complete. Particularly this is assumed to be an issue for the full-text corpus since it had the least favorable coverage.

Besides time, an important variable in the analyses is locality, which is represented by the five major universities, each with its own department of sociology, producing dissertations in sociology in Sweden. The dissertation production of these universities as well as four smaller universities and university colleges given the right to award PhDs in sociology, represented by the category Other, is illustrated in Figure 33.

![Graph of dissertation production by university](image)

**Figure 33.** The production of dissertations in sociology in Sweden by time and place. Absolute numbers.

In Figure 33, it is shown that three universities – Uppsala, Stockholm, and Lund – were the sole producers of dissertations in sociology from the 1950s to mid-1960s, when they were adjoined by two other universities – Gothenburg followed by Umeå. Later, in the early 2000s, the rest of the universities and university colleges also began to print dissertations. Further, the production is not evenly distributed and the universities have different periods
where they produce more or less of the average. For instance, Lund University represents the largest overall producer, adjoined by Uppsala University in the 20th century and Stockholm University in the 21st century.

Figure 34. Local distribution of the two corpora in relation to the bibliography (physical books). Absolute numbers.

Figure 34 is a bar chart showing the local distribution of dissertations in sociology compared to the corpora that are prepared for analysis, which gives us four variables: on the one hand physical books and abstracts produced 1980-2019, and on the other hand physical books and full-texts produced 2000-2019. Lund University represents the largest share of physical books for the period 1980-2019, followed by Stockholm University and then Uppsala University. For the rest of the universities, Gothenburg University made more books in that period than Umeå University, which, in turn, produced almost twice as many as the rest of the universities and university colleges together. For the most part, this pattern is echoed in the abstract corpus that has been prepared for analysis, which has only a few missing values. The exception is found in Lund University which has a lower abstract/physical book ratio compared to the others. The reason for this is simply that dissertations in sociology from that university were not always coupled with an abstract in the earlier part of the period (1980-2000). Thus, it is worth keeping in mind that one setback with the abstract corpus is the
potential risk that the full richness of the dissertations produced at Lund University is not represented.

Turning to the other dataset, the full-text corpus, Figure 34 shows that the main universities have a comparatively equal share of the full-text corpus. However, Stockholm University diverges from this pattern with its relatively large number of full-texts (n=112) compared to the other four main universities (n=49-67). Further, it is worth acknowledging the full-text/physical book ratio. Here Lund University and Gothenburg University have a bit weaker ratio than the other universities in the sample whereas Umeå University have the strongest ratio. When expanding on this problem beyond what is picked up in Figure 34, it was found that all universities have good coverage for the second part of the period (2010-2019), but the same does not hold true for the coverage of the first part of the period (2000-2009). For the whole period (2000-2010), Umed University stands out with a 96 percent coverage, while Lund University and Gothenburg University score below average with about 65 percent.

The fact that the sample is not complete should be kept in mind when interpreting the results. However, it is difficult to speculate if a more thorough coverage would have entailed major differences in the general structure of models generated on the global level of the corpora. This includes whether the overarching character of the most central word constellations, topics, linguistic styles, and word embeddings would have been altered or not. It is possible that the divergence in coverage between some periods and some universities could have affected the interpretation of the conditioning of sociological knowledge. Since the comparisons are based on relative measurements the skewed coverage would primarily affect the global level of the corpus, which, nonetheless, conditions the temporal and spatial distributions.

This appraisal of the abstract corpus and the full-text corpus has suggested a few limitations. In general, the corpora have a better representation in the later part of their respective period. Particularly the early period of the full-text corpus (2000-2009) is problematic and might influence how the results ought to be interpreted. Further, some localities, like Lund University, are a bit under-represented in the corpora in relation to their actual production of dissertations in sociology, whereas others, like Stockholm University and Umed University, are a tad over-represented. However, given the nature of the methods, where many methods are based on stochastic processes, it is particularly hard to foresee the potential consequences of these biases as there are already so many parameters to be set that alter the outcome in various ways. Further, while these biases will certainly affect the global level (although it is hard to predict precisely how) all analyses of the social conditioning of knowledge are based on relative measurements.
Appendix C: Topic Modeling

Topic modeling is the primary text mining method applied in Chapter 7. In basic terms, topic modeling is about discovering thematic structures in corpora, annotating the documents constituting the corpora, and using the annotation for visualizations. Although it originally developed within computer science, topic modeling has started to gain momentum in the social and human sciences due to its ability to inductively process large corpora. According to Mohr and Bogdanov (2013, p. 546), one of the expedient aspects of, topic modeling is that it is based on the old Saussurean axiom that linguistic entities get their meanings based on their relation to other entities within a system. Topic modeling “enable us to take the measure of large-scale social phenomena that we could not have previously been able to do” (Mohr and Bogdanov, 2013, p. 561).

The particular topic model applied in Chapter 7 is the most popular one, *Latent Dirichlet Allocation* (LDA). LDA is a generative probabilistic model of a corpus that gives “latent variables […] representing probability distributions on sets of words” (Blei et al., 2003, p. 996). In LDA, a corpus is assumed to be constituted of documents that are distributions over topics, where a topic is a distribution over words. To put it differently, a topic is a latent variable constituted by words – or, maybe more precisely, a low-dimensional set of multinomial distributions – that the algorithm finds to recur within documents.

LDA is based on a Natural Language Processing model referred to as Bag-of-Words (BOW), in this case with a unigram type of n-grams. The LDA BOW entails that documents are broken down to collections of tokens or single words referred to as unigrams. The presented order of the tokens is irrelevant, hence the BOW metaphor, which enables the creation of vectors (i.e., probabilistic distributions) to compare the frequencies of tokens over documents. In other words, “topic models capture co-occurrences regardless of these words’ embeddedness within other complexities of language” (Mohr

---

163 A unigram is a token of one word or set of connected letters without spaces (e.g., “BOW” and “Bag-of-Words”). In comparison, a bigram consists of two words occurring together (e.g., “text mining” and “qualitative methods”) and general n-grams take into consideration expressions of an unlimited number of words that re-occur side by side throughout the text.
and Bogdanov, 2013, p. 547). Thereafter, all documents making up the corpora are transposed into a Document-Term-Matrix (DTM), where each document makes up a row and every token or unigram corresponds to a column.

Thereafter, the LDA can in principle be performed, but as was elaborated on in the section above, documents often have to be pre-processed before proceeding. The overall steps in the standard sequence are removing punctuations, transposing all words to lowercase, morphing words into stems, removing numbers, and adding a standard vocabulary of *stop words* (e.g., “the”, “a”, “in” etc.) as well as remove the most frequent terms (e.g., “also”, “can”, “differ”, “one”, “relat”, “research”, “social”, “studt”, and “use”) for allowing terms with more valuable semantics to constitute the topics.

An instructive presentation of how the algorithm functions can be found in the original paper on LDA (Blei et al., 2003), which is presented in Figure 35. In the center of it, we find “the word” level of all documents, \( N \), and within it a representation of repeated occurrences of topics, \( z \), and words, \( w \) (sampled once for every term). Outside \( N \) we find \( M \), which is “the document level”, and within it a representation of the mixture of topics, \( \theta \) (which is sampled once for each document). Finally, outside \( N \) and \( M \), is the corpus level, where a representation of the word vector, \( \alpha \), and topic \( \times \) word matrix, \( \beta \), lies (each sampled at one occasion only). In practice, the \( \beta \) are the most important outputs since they give an order to the topics by presenting probabilities for each term to occur within their assigned topics. For every topic, its meaning is given by interpreting the relationship between the words assigned to it – in particular the words with the highest \( \beta \)-value.

![Graphical model representation of LDA (Blei et al., 2003, p. 997)](image)

One central assumption in almost all methods for finding latent topics is that the number of topics in a corpus, \( k \), is assumed to be fixed and known beforehand by the analyst. Thus, since the model does not provide a posteriori number of topics, the analyst is forced to determine how many topics it should generate. There are, however, a few unsupervised methods to get a sounder selection, mathematically speaking, and, in this case, to avoid having to run hundreds of test runs on corpora as large as 10,000 pages. The
four most popular are the *Latent Concept Modeling* (Deveaud et al., 2014), the symmetric Kullback-Leibler divergence (Arun et al., 2010) the density-based method (Cao et al., 2009), and the Markov chain Monte Carlo method (Griffiths and Steyvers, 2004).

![Held–Out Likelihood](image1)

![Residuals](image2)

![Semantic Coherence](image3)

![Lower Bound](image4)

Figure 36. Diagnostic values supporting the choice of K (number of topics) for the structural topic model fitted to the abstract corpus.

A selection of typical evaluation metrics is presented in Figure 36. The aim is here to get high held-out likelihood and semantic coherence and low residuals and lower bound. For every topic modeling fitted to the abstract corpus, all four metrics were run and compared to find the ultimate fit. After also adding qualitative evaluations of the different number of topics (ranging from 15 to 35), 20 topics were found. Thereafter several models with 20 topics were run and evaluated with a second algorithm, comparing the trade-off between the semantic coherence and the exclusivity of the topics. After the topic modeling was run to extract the topics from the corpus, a qualita-
tive reading of the topics was conducted to generate the labels. This process is thoroughly explained in Chapter 7.

In Chapter 7, a variant of LDA is applied called the structural topic model or STM (cf. Roberts et al., 2019, 2014). STM has become popular in the social sciences since it allows for topics to be a function of the metadata of documents. In other words, it enables to estimate the effect of document covariates on the prevalence and content of topics. This is done through a logistic-normal generalized linear model that can measure the document-specific distribution of the words constituting each topic. Within the frame of this dissertation, STM is utilized to investigate how the place where and the year when the dissertations were produced might affect what topic is selected.

It is worth mentioning that topic modeling is based on stochastic processes. This entails that they can be applied to predict overarching patterns and trends, but that the final product will differ every time you train the model (depending on where you set the seed that is). Focusing on the case of topic modeling, the complete sample of a given corpus does not equate to generating the best possible model. Quite the opposite, recent developments in the field of computational social science have suggested that it is better to leave out a good share of the corpus for evaluating and finetuning the topic model (cf. Stewart et al., 2022). For instance, when performing this form of unsupervised machine learning on text data, the corpus ought to be split into a training set (e.g., 80% of the corpus), on which the model is trained, and a test set (e.g., 20% of the corpus), on which the performance of the trained model is tested. However, at the time of training the structural topic model (mid-2021), my reading of the developments of the computational text analysis field in sociology was that the common practice was to model on the complete corpus. This can then be seen as a potential limitation of this particular model.
Appendix D: Stylometry

Appendix D deals with the technicalities of the methods applied in Chapter 8, which is dedicated to generating and interpreting linguistic styles from dissertations in sociology. Linguistic style is a concept derived from the field of stylistics that entails a manner of using language. Within stylometry, a branch of applied stylistics intersecting with fields like computational linguistics and digital humanities, the concept of linguistic style has a stricter function. Here linguistic styles simply mean feature variations or, in practical terms for the case of written text, patterns of term frequencies based on word occurrences in the documents of study. While many methods rely on term frequencies, there is a fundamental difference between stylometry and, for instance, the topic modeling process described in Appendix B: stylometry targets the most frequently occurring words (MFWs) to detect stylistic patterns, rather than less frequently occurring words that topic modeling utilizes to identify topical patterns. The stylistic patterns modeled by stylometry are constituted by word repetitions that are often undetectable to the human eye, which is a reason why stylometry is often applied in author attribution 164 and in Chapter 8 to detect formalized ways of writing in sociology that exists “underneath” the level of topics. As one can imagine, finding such latent patterns of word repetitions requires a large number of text samples, which is why trained on the full-text corpus of dissertations in sociology.

This appendix will start by presenting the reasoning behind choosing the bootstrapped consensus trees to compute linguistic styles, which is followed by a description of the architecture and main intuition behind this method. Thereafter, the evaluation process and parameters of the network presented in the analysis are given, which is trained on the full-text corpus of dissertations in sociology. Further, the appendix will describe the modularity algorithm leading to the five linguistic styles presented in the analytical chapter as well as how these styles were interpreted through their Zeta values and close readings.

164 Author attribution is a particular research problem where the goal is to find the author of an unnamed document, which is often solved by comparing its content to authored documents with similar linguistic styles.
One experience-based assumption behind the current state of the art in stylometry, formulated by a key figure named Burrows (2002, p. 268), is that “...a wealth of variables, many of which may be weak discriminators, almost always offer more tenable results than a small number of strong ones.” The algorithm applied here is bootstrapped consensus trees where “the bootstrap procedure will run different (virtual) cluster analyses and aggregate the results into a single (unrooted) consensus tree” that can be visualized as a single consensus network (Eder et al., 2016, p. 117)

When measuring the Delta or distance between documents by computing difference in word occurrences, we are dealing with high-dimensional data. In the statistical toolbox, there are here a few “natural” tools to handle with this problem. Assuming that we want to measure the distance between $u$ and $v$, one tool would be the Manhattan distance ($\sum_i |u_i - v_i|$) and another the Euclidean distance ($\sqrt{\sum_i (u_i - v_i)^2}$). However, this analysis will apply another Delta distance known as is Burrow’s Delta (cf. Burrows, 2002) – described as $\Delta_B = \sum_{i=1}^n |z_i(D_1) - z_i(D_2)|$, where (cf. Evert et al., 2017) – since it is more commonly applied in computational linguistics for detecting stylistic differences and authorship within.\footnote{Another valid option would have been to compute the Delta distance based on cosine similarity, a common algorithm in text mining, yet it was decided to go with the most vetted method in the field of stylometry (cf. Evert et al., 2017).}

First, the distances between documents are measured with a version of Euclidean (normalized) distance known as the Burrow’s Delta (cf. Burrows, 2002), which is often used. Second, Ward’s linking algorithm (cf. Ward, 1963) is applied to form hierarchical groups from mutually exclusive subsets with maximal similarity in some specified characteristics. The specific procedure performed is culling corpora from 20 to 80 with an increment of 20 (i.e., 20, 40, 60, and 80). For each culled corpus, the most frequent words (MFW) for the set of documents are computed. The model takes 50 MFW to 1550 MFW with increments of 50 (i.e., 50, 100, 150, 200 ... 1550), and the cutoff is 5000 unique words. The parameters are set to a consensus strength of 0.5 and is based on words (unigrams).

Through Gephi, the task of “extracting the strongest patterns... and filtering out weaker textual similarities” (Eder, 2017, p. 56) is followed up by only keeping the strong links. Community detection is made by measuring the modularity of the weighted graph with the Louvain method (cf. Blondel et al., 2008), which can be defined as
\[ Q = \frac{1}{2m} \sum_{ij} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \]

where \( A_{ij} \) is the edge weight between the \( i \) node and the \( j \) node, the edge weights of these nodes are denoted by \( k_i \) and \( k_j \), \( m \) holds the sum of all edge weights in the network, \( c_i \) and \( c_j \) represents the node communities, and \( \delta \) represents the Kronecker delta function.

With a set of communities at hand that depict the overall linguistic style of each full-text, the next step is to investigate the most discriminative stylistic features of each community. This is done by opposing them to each other by comparing their Craig’s Zeta scores (cf. Craig and Kinney, 2009), an extension of Burrow’s Zeta (cf. Burrows, 2007). This operation is conducted in binary classification settings where lists of tokens are generated that are constituted by the words that are most typical and least typical for each community vis-à-vis the other styles. The most discriminative words for each linguistic style community are then visualized through a network with the words as nodes and the edges of the Zeta scores.

One clear limitation of this modularity algorithm is that a node can only belong to one community, which is restrictive to borderline cases that lie in between two or more linguistic styles. To counter this problem, the analysis is supplemented by qualitative content analysis based on close readings that only consider the most central dissertations in each community. These are computed through the bootstrapped consensus network by taking the nodes with the highest ratio of edges within the community when edges shared with vertices outside the community are subtracted.
Appendix E: Word Embeddings

This appendix deals with the technical aspects of the methods applied in Chapter 9, which exclusively focuses on training and analyzing word embeddings through shallow neural networks. The concept of word embedding is here used in the sense of a distributional representation of a term in the form of vectors. Together, the word embeddings form a semantic vector space where close proximity between word embeddings is expected to entail semantic similarity. One of the key benefits of word embeddings over other word vector approaches is that they are able to capture word occurrences within a k-dimensional space while preserving the local context of the words by representing their relations in the form of distances. The goal of Chapter 8 is to outline complex meaning of words by mapping how they are used in the corpus, which coincides well with the overarching idea of word embeddings to model distributional semantics over a vector space. The structure of the appendix is as follows. First, the reasoning behind selecting the continuous skip-gram model from word2vec is offered. Thereafter, a description of the architecture and main intuition behind the model is given. Lastly, the stats and parameters of the neural network model utilized in the analysis are presented, which is trained on the full-text corpus of dissertations in sociology.

In recent years, the dissemination and application of word embeddings have severely intensified in general society in the form of pre-trained Large Language Models (LLMs) like ChatGPT4. However, the idea of probabilistic modeling of language through neural networks has circulated in academia for at least two decades (cf. Bengio et al., 2003). For the last decade, there have been a few popular frameworks for generating word embeddings from a local corpus (i.e., not pre-trained word embeddings like in ChatGPT4). When researching the literature, it was found that the most commonly used in social science were Word2vec (Mikolov et al., 2013), either the Continuous Bag-Of-Words model (CBOW) or the Continuous Skip-gram model, GloVe or Global Vectors for Word Representation (Pennington et al., 2014), and BERT or Bidirectional Encoder Representations from Transformers (Devlin et al., 2018).
Following this observation, all four models mentioned were trained on the full-text corpus. After conducting common tasks like testing a set of relevant word analogies and semantic similarities, a comparison was made for the results of each model. On this basis, a decision was to go with Word2vec’s Continuous Skip-gram model, since it performed best on the tasks executed on this relatively small dataset. All analyses were performed in the statistical programming language R, particularly the packages word2vec, text2vec, text2map, igraph, tm, and tidytext.

The Word2vec is a so-called neural probabilistic language model (cf. Bengio et al., 2003) that utilizes neural networks “for learning the distributed representations of words” in the form of continuous vectors (Mikolov et al., 2013, p. 4). The original paper by Mikolov and colleagues (2013) presents two log-linear models. The intuition behind the first model, the CBOW, is to predict a target word given a set of context words, whereas the second, the continuous skip-gram model, instead seeks to predict the context words given a target word. Since the goal of the study of word embeddings presented in this dissertation is to find the distributional representation of specific target words like sociology and society, the latter model appears to be a better fit.

In the more technical descriptions of the original paper (Mikolov et al., 2013, p. 4), training the continuous skip-gram model is proportional to

\[ Q = C \times (D + D \times \log_2(V)) \]

where \( Q \) represents the algorithm’s computational complexity, \( C \) is the maximum distance of the words (i.e., context or word-window size), \( D \) captures the word representations, and \( V \) symbolizes the size of the vocabulary (i.e., the given dictionary of unique words) that is simplified through the

These parameters decide the borders of the continuous vector space modeled by the algorithm, which is constituted by vectorial representations of terms (i.e., word embeddings). Based on this setup the model can “predict words within a certain range before and after the current word” (Mikolov et al., 2013, p. 4) and the architecture for how this procedure is performed through a neural network can be graphically illustrated like this (Figure 37)
where the target word $w(t)$ is inserted in the INPUT layer, the PROJECTION is a hidden layer constituted by the dot product between the $w(t)$ and the weights, and the OUTPUT layer gives the dot product between the projection (i.e., the output vector of the hidden layer) and the weights of the output layer. Given a context represented by a word window of two, the model attempts to maximize the probability of observing the four context words – by $w(t-2), w(t-1), w(t+1),$ and $w(t+2)$ – to the target word $w(t)$.

Again, the aim of the continuous skip-gram model is to predict probabilities of context words that are as close as possible to the actual context word probabilities in the corpus. To convert the similarities (dot products or inner products) into probabilities and compute the probability distribution over context words for the target word, the softmax function is utilized (cf. Bridle, 1989). In brief terms, the function converts a vector of $K$ real numbers into a probability distribution of $K$ possible outcomes as of this formula

$$
\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, 2, \ldots, K
$$

where the softmax is represented by the sigma sign, $\sigma$, $K$ is the number of classes in the multi-class classifier, $e^{z_i}$ represents the standard exponential function for the input vector and $e^{z_j}$ the standard exponential function for the output vector.
The softmax function is used to predict the discrete multinominal probability distributions of the context words based on the dot products between the vector of the target word and the vectors of the context words. To unpack, a context word with a probability score of 1.00 to the target word entails that the two word embeddings occupy the same position in vector space. Conversely, a score of -1.00 would entail that the words have diametrically opposite positions in vector space. The parameters of the model are then updated iteratively and optimized through stochastic gradient descent and backpropagation (cf. Rumelhart et al., 1986), which is a commonly used weight-adjusting procedure in machine learning. This technique is utilized to make the predicted probabilities closer to the true probabilities in the training data or, to put it differently, to minimize the difference between the actual and the desired output vector by generating ‘hidden’ features able to capture regularities.

When generating word embeddings from local data, a few decisions have to be made and parameters set. The original corpus is made up of almost 23 million words distributed over 380 sociology dissertations in full text. After observing the most infrequent words in the corpus, it was found that typos and incomplete words had to be cleaned out to remove unnecessary noise that would compromise the performance of the model. A decision was made to only include words occurring at least three times in the corpus, leaving us with 66,000 unique words.

Thereafter, the size of the word window had to be selected, that is how many context words before and after the target word ought to be included in the model. In comparisons of different word window sizes for the continuous skip-gram model (cf. Levy and Goldberg, 2014), it was found that smaller windows (n=2) are good for computing similar words to the target word and larger windows (n=5) are better at capturing domain-specific information like what type of words occur in related discussions. Since the task of the analysis is to explore in what context words like sociology and society are used in the corpus, the natural choice was a larger word window. After experimenting with word window sizes between five and ten, it was decided that eight was the best size. This leaves us with symmetrical context to the target word $w$ constituted by the eight context words positioned before ($w^8$, $w^7$, $w^6$, $w^5$, $w^4$, $w^3$, $w^2$, and $w^1$) and the eight context words positioned after ($w^1$, $w^2$, $w^3$, $w^4$, $w^5$, $w^6$, $w^7$, and $w^8$).

One issue that is more specific for word embeddings is how many dimensions the vector space ought to be constituted by, that is how many vectors will each word embedding be represented by. This choice comes with a few different trade-offs, like that fewer dimensions risk to reduce the functionality of the model or will not be able to capture enough nuances in its predic-
tions and that adding additional dimensions makes the computation more costly. Following the pre-trained models presented in key publications such as the Wiki News and the Common Crawl (cf. Mikolov et al., 2017; Pennington et al., 2014), 300 dimensions appear to be the current golden standard. Since this study is made on a smaller corpus or dataset, several different numbers of dimensions, spanning from 50 to 500, were tested before finally settling on 300.

The final model was run with 20 iterations and consists of 66,000 word embeddings with 300 dimensions, which gives us a vector space constituted by almost 20,000,000 elements. A simplified representation of this space can be found in Figure 38.

![Figure 38](image.png)

Figure 38. The vector space constituted by skip-gram word2vec word embeddings modeled on the sociology dissertation full-text corpus.

Figure 38 is a dimensionality reduction of the continuous skip-gram model based on the full-text corpus computed with the Uniform Manifold Approximation and Projection (UMAP) method with cosine similarity as the chosen metric (cf. McInnes et al., 2020). Each gray dot here represents a word embedding and proximity entails similarity. While this figure is only presented
for the reader to get an idea of how thousands of words can be converted into sequences of numbers and combined to form a vector space, we can sense that the words seem to form clusters of words that are relatively well spread over these two dimensions.

Following the distributional hypothesis – which is often popularly summarized in the phrase “You shall know a word by the company it keeps!” (Firth, 1962, p. 11) – closeness in vector space is hypothesized as equivalent to closeness or remoteness in semantic similarity. In the case of word embeddings, cosine similarity is used to compute the probability of context words given a target word (e.g., Jatnika et al., 2019; Sitikhu et al., 2019). In simple terms, cosine similarity can be defined as the dot product of two vectors divided by the product of their magnitudes, which corresponds to the cosine of the angle between them. In the analysis, cosine similarity is, for instance, utilized to predict the most probable context words to occur next sociology, which followed then taking these words as target words and computing their context words. When converting the words to vertices and the probabilities to edges, we end up with a network for interpreting the distributional meaning of sociology.

Besides predicting similarities of context words occurring next to a target word, one can also calculate the distance between documents and word embeddings in the modeled vector space (cf. Kusner et al., 2015). While comparing word for word in all documents would be demanding, a less computationally costly method to measure cosine distance was presented under the name the Relaxed Word Mover’s Distance or RWMD (cf. Stoltz and Taylor, 2019). With the goal to predict what documents engage with a chosen concept (represented by one or a few target words), RWMD was further developed by sociologists into the Concept Mover’s Distance or CMD (cf. Stoltz and Taylor, 2019). CMD is put into practice by developing an algorithm that computes “the minimum cost that a document’s embedded words need to travel to arrive at the position of all the words in an ideal ’pseudo document’ consisting of only words denoting a specified concept” (Stolz and Taylor 2019, p. 294).

CMD is repeatably utilized on the vector space to find out the extent to which the sociology dissertations analyzed engage with different concepts and how this engagement differs in relation to temporality, spatiality, and linguistic style by sorting the corpus in to sub-corpora (i.e., year of publication, university where the defense took place, and the dissertation’s assigned linguistic style). In both cases – the word embeddings’ closeness to each other and to documents respectively – the distance metric is based on cosine similarity.
42. Ulla Jergeby: Offentlig miljö som arena och kuliss. Att se, bli sedd och mötas på stadens offentliga och bostadsnära platser. 1996.
47. Peter Sohlberg: Mål och mening i samhället. Funktionalistiska program i samhällsvetenskapen. 1999.
63. Fereshteh Ahmadi (ed.): Coping with Cancer in Sweden – A Search for Meaning. 2015.