



# ANALYZING MEANING IN BIG DATA: PERFORMING A MAP ANALYSIS USING GRAMMATICAL PARSING AND TOPIC MODELING

**Jan Goldenstein\***

**Philipp Poschmann\*** 

## Abstract

*Social scientists have recently started discussing the utilization of text-mining tools as being fruitful for scaling inductively grounded close reading. We aim to progress in this direction and provide a contemporary contribution to the literature. By focusing on map analysis, we demonstrate the potential of text-mining tools for text analysis that approaches inductive but still formal in-depth analysis. We propose that a combination of text-mining tools addressing different layers of meaning facilitates a closer analysis of the dynamics of manifest and latent meanings than is currently acknowledged. To illustrate our approach, we combine grammatical parsing and topic modeling to operationalize communication structures within sentences and the semantic surroundings of these communication structures. We use a reliable and downloadable software application to analyze the dynamic interlacement of two layers of meaning over time. We do so by analyzing 15,371 newspaper articles on corporate responsibility published in the United States from 1950 to 2013.*

---

\*Friedrich Schiller University Jena, Thuringia, Germany

## Corresponding Author:

Philipp Poschmann, School of Economics and Business Administration, Friedrich Schiller University Jena, Carl-Zeiss-Straße 3, Jena, Thuringia D-07743, Germany.

Email: [philipp.poschmann@uni-jena.de](mailto:philipp.poschmann@uni-jena.de)

**Keywords**

*text analysis, natural language processing, grammatical parsing, topic modeling, corporate responsibility*

**1. INTRODUCTION**

The recent trend toward big data promises that instead of being restricted to studying comparably small pieces of information, social scientists can consider large amounts of text in their full complexity and nuance. Recently introduced methodological frameworks consider the utilization of contemporary text-mining tools as fruitful for scaling inductively grounded close reading (Edelmann and Mohr 2018; Mohr, Wagner-Pacifici, and Breiger 2015; Muller et al. 2016; Nelson 2017) while retaining an inductive in-depth analysis of the complex layeredness and the dynamics of manifest and latent meanings in texts (Mohr et al. 2013, 2015; Popping 2012; Vicari 2010). These methodological frameworks either focus on one layer of meaning by relying solely on a specific text-mining tool to detect coarse-grained meaning patterns (DiMaggio, Nag, and Blei 2013; Jockers and Mimno 2013; Rule, Cointet, and Bearman 2015) or they combine a text-mining tool with close reading. In the latter, close reading is used to validate the results provided by text-mining tools or support inductive in-depth analyses by focusing on selected meaning patterns (Abramson et al. 2018; Breiger, Wagner-Pacifici, and Mohr 2018; Chakrabarti and Frye 2017).

We draw on the essence of these methodological frameworks and propose their extension by considering text-mining tools' potential to further approach inductive but still formal in-depth analysis. We propose a middle way between the exclusive utilization of text-mining tools and their combination with close reading. We argue that the combination of text-mining tools that address different layers of meaning has the potential to facilitate a closer analysis of the dynamics of manifest and latent meanings than is currently acknowledged. That is, the formal in-depth analysis of the dynamic interlacement of layers of meaning opens space for more contextualized and fine-grained interpretations of texts (Mohr et al. 2015). To be sure, we do not claim that close reading can or should be replaced by text-mining tools. On the contrary, our approach highlights that utilization of text-mining tools facilitates a formal in-depth analysis, enabling first interpretations that in turn may provide a stronger focus for subsequent close reading.

To apply text-mining tools for an inductive but still formal in-depth analysis, we draw on an ongoing discussion in the social sciences about the transformation of large text corpora into maps. According to Lee and Martin (2015), a map is a focused representation of the *entire* text corpus and depicts the relationship of textual characteristics, such as parts of speech, the grammatical connection of words, or latent semantic patterns. In this regard, maps project complex relationships into a reduced *space* and can provide information on the proximity or connectedness and the scale or proportion of textual characteristics. We use different text-mining tools to demonstrate that the transformation of manifest and latent meanings into maps is especially well suited for analysis of the dynamic interlacement of different layers of meaning. To this end, we consider the first layer to be the meaning, which is constructed from communication structures within sentences. Communication structures depict semantic triplets, namely, the grammatical connection of subjects, verbs, and objects that uncover manifest narratives of social action in and across time (Franzosi 1989; Roberts 1989). A second layer of meaning acknowledges that sentences are embedded in latent semantic surroundings that significantly influence their actual meaning (Kahl and Grodal 2016; Khaire and Wadhwani 2010; Kirchner and Mohr 2010; Wagner-Pacifci, Mohr, and Breiger 2015). This means words and sentences do not carry one static meaning; rather, their meaning is dynamic and influenced by the semantic surroundings they are embedded in, and vice versa (Krippendorff 2004; Popping 2012; Vossen 2004).

In this article, we illustrate the application of our methodological framework using two different text-mining tools for map analysis: grammatical parsing and topic modeling (for an overview of other text-mining tools that address different layers of meaning, see Evans and Aceves 2016). We use grammatical parsing to operationalize communication structures in sentences (e.g., Carroll 2004), and we build on topic modeling (e.g., Blei, Ng, and Jordan 2003) to operationalize the semantic surroundings of these communication structures. Both kinds of text-mining tools have been attributed high relevance in the social sciences (Cornelissen and Werner 2014; Evans and Aceves 2016; Goldenstein et al. 2019; Mohr and Bogdanov 2013; Sudhahar et al. 2013; van Atteveldt, Kleinnijenhuis, and Ruigrok 2008). However, the combined utilization of text-mining tools to transform a complete text corpus into maps to analyze the dynamic interlacement of different layers of

meanings depicts a novel endeavor. To the best of our knowledge, the only analysis that uses text-mining tools to extract different layers of meaning was conducted by Mohr and colleagues (2013). Yet, in line with existing methodological frameworks, Mohr and colleagues did not use text-mining tools for a formal in-depth analysis. For their co-occurrence analysis, they only used parts of their text corpus, did not consider explicit communication structures, and did not explore the dynamic interlacement of these communication structures with their semantic surroundings over time. Our methodological approach is therefore a valuable contribution to contemporary frameworks that use sophisticated text-mining tools.

In the following sections, we expound our methodological contribution. We start with an overview of how formal text analysis has been used in the social sciences so far, and we highlight the recent trend toward big data analysis using text-mining tools. Next, we outline our methodological framework and focus on its application using grammatical parsing and topic modeling. To make the combination of these text-mining tools accessible for and applicable to social scientists, we provide a reliable and accurate software application.<sup>1</sup> We apply our framework to a text corpus of 15,371 U.S. newspaper articles. These newspaper articles cover the responsibilities attributed to companies and their representatives in the United States between 1950 and 2013. Finally, we discuss the benefits of our framework for social science research and point to potential areas for further development.

## **2. AN OVERVIEW ON FORMAL TEXT ANALYSIS IN THE SOCIAL SCIENCES**

### *2.1. Coding: From Deductive Content Analysis to Structural Approaches*

Deductive content analysis, from the start, focused on the measurement and transformation of the occurrence and distribution of linguistic units into numeric variables. In an early attempt, Lasswell, Lerner, and Pool (1952) coded predefined keywords, phrases, and categories, a procedure that has, in essence, been adopted unaltered, even by more recent work (Abrahamson and Eisenman 2008; Palmer, Kabanoff, and Dunford 1997; Short, Broberg, and Brigham 2010). Deductive content analysis has been criticized, however, for neglecting the fact that words are

meaningfully related to one another (Pool 1959). Quantitative co-occurrence analysis was thus created to consider the relationships between keywords, phrases, and categories (Kabanoff, Waldersee, and Cohen 1995; Kennedy 2008). However, co-occurrence analysis does not reveal the grammatical connection between words as it only explores word spans; therefore, at best, it only functions as a proxy for grammatical relationships. To address this challenge, scholars created semantic grammar (Abell 1987), which is a hand-coding approach to capture and count semantic triplets—namely, the grammatical connection of subjects, verbs, and objects—to construct quantitative variables (Franzosi 1989, 1990; Roberts 1989, 2000).

Since at least the 1990s, social scientists have explored new methodological ways to enable the measurement of more latent meaning, that is, meaning that exists below the surface of texts and requires consideration of their inherent complexity (Mohr 1998; Mohr et al. 2015). Instead of counting the occurrence, co-occurrence, or grammatical connection of linguistic units, this kind of analysis focuses on the structured representation of specific textual content (Nelson 2017). For example, structural approaches construct networks (Carley 1997; Franzosi, de Fazio, and Vicari 2012) or multidimensional scaling plots (Meyer and Höllerer 2010) and use the strength of the relational association of the previously defined linguistic units to position these units in two-dimensional spaces. The invention of these structural approaches was an important step toward a formal text analysis that does not distort its inductive character.

However, due to technical limitations, structural approaches inherited some traditions from deductive content analysis, namely, *ex ante* decisions about which linguistic units should be considered and how the strength of their association should be measured (Carley 1993; Mohr 1998). As Biernacki (2012) notes, *ex ante* interpretations are problematic because they involve the necessarily subjectively driven exclusion of linguistic units, or the grouping of particularities into labeled categories beyond the observer's sight. As such, coding "cannot take one beyond the confines of what has been generally supposed at the outset" (Biernacki 2009:179). In addition, Lee and Martin (2015:4) state that coding hardly "allow[s] us to understand that the part is only interpreted by reference to the whole, the part is fixed, determined, stapled to its interpretation via the violence of 'proof by assertion.'"

Lee and Martin (2015) argue that counting (rather than coding) avoids the problems associated with ex ante interpretations. They suggest counting textual characteristics that do not require any ex ante interpretative act but instead support inductive in-depth analyses by focusing subsequent close reading. In other words, the results of a textual analysis should emerge through uncovering patterns of meaning in the data, should not be influenced by the coding procedure applied, and should be open for undistorted in-depth analysis. One way to follow Lee and Martin's (2015) suggestion is, in light of the arrival of big data, to use sophisticated text-mining tools.

## *2.2. Counting or Focusing? Text-Mining Tools and In-Depth Analysis*

Recently introduced methodological frameworks have suggested the exclusive utilization of text-mining tools to count textual characteristics without ex ante interpretations (Mohr et al. 2015). For example, by using stylometry software, social scientists have suggested counting the most frequent style markers or morphological features to analyze style variations in texts on a large scale (Allison et al. 2011; Eder 2014; Moretti 2013). Rule and colleagues (2015) used a phrase parser to extract noun phrases from State of the Union addresses given from 1790 to 2014. They counted the co-occurrence of these noun phrases to represent this discourse and inductively identify how usage of terms and categories changed over time. Sudhahar and colleagues (2013) drew on the idea of semantic triplets and used a grammatical parser to extract some predefined triplets from public discourse. They coded verbs to use them as edges in a network that maps whether the relationship between persons (subjects) and specific nouns (objects) is positive or negative. However, by preselecting triplets and predefining verbs as negative or positive, Sudhahar and colleagues still followed ex ante coding decisions. Some researchers have applied topic modeling to their text corpora. For instance, Jockers and Mimno (2013) studied the correlation of external factors, such as author's gender, author's nationality, and date of publication, with the choice of literary topics in novels. DiMaggio and colleagues (2013) investigated how government assistance for artists and arts organizations has been framed in different arenas of public discourse.

By relying exclusively on one text-mining tool, these attempts have been criticized for taking the results of text-mining tools as given and neglecting the inductive character that in-depth analysis based on close reading could provide (Breiger et al. 2018; Muller et al. 2016). Consequently, social scientists have developed methodological frameworks that use one text-mining tool combined with close reading. Under these frameworks, text-mining tools are first applied to conduct an analysis that uncovers coarse-grained patterns of meaning. Focusing on selected meaning patterns, close reading is then used for inductive in-depth analysis. As a result, inductive text analysis should become more formalized, transparent, and replicable (Abramson et al. 2018; Baumer et al. 2017; Chakrabarti and Frye 2017; Muller et al. 2016; Nelson 2017).

However, methodological frameworks that rely exclusively on text-mining tools or that combine a text-mining tool with close reading neglect text-mining tools' potential for approaching an inductive but still formal in-depth analysis. We argue that text-mining tools already have the potential to facilitate a closer analysis of the dynamics of manifest and latent meanings than is currently acknowledged (Breiger et al. 2018). We thus propose a middle way that uses a combination of text-mining tools to conduct a formal in-depth analysis, enabling first interpretations that in turn may support a stronger focus for subsequent close reading analyses. To apply the combination of text-mining tools for such a formal in-depth analysis, we suggest transforming large text corpora into *maps* (Lee and Martin 2015).

### *2.3. Map Analysis: Counting and Mapping the Layeredness and Dynamics of Meaning*

Based on Lee and Martin's (2015) work, maps exhibit the following features: (1) A map is based on counting textual characteristics without ex ante interpretations, and (2) a map depicts a focused version of the entire text corpus from the perspective of the chosen textual characteristics. In doing so, maps can provide information about the proximity or connectedness and the scale or proportion of textual characteristics. We add the following: By combining different textual characteristics that represent different layers of meaning, such as parts of speech, the grammatical association of words, or the existence of latent semantic patterns, maps

can also depict the complex layeredness and dynamics of meaning in a text corpus.

Map analysis can thus be meaningfully supported by modern text-mining tools that are well equipped to count the occurrence of textual characteristics on a large scale. Map analysis enables these characteristics to be projected into a reduced map space and provides a focused perspective on specific meanings hidden in large text corpora while leaving others intentionally unconsidered. According to Lee and Martin (2015), the usefulness of maps is especially due to their focus on different textual characteristics that help highlight multiple perspectives on texts. This kind of focus does not—in contrast to a coding procedure that requires *ex ante* interpretations—force any particular meaning on texts. Thus, maps enable first and undistorted *ex post* interpretations that may support a stronger focus for subsequent close-reading analyses.

In this article, we take a novel approach to demonstrate how text-mining tools enable the transformation of manifest and latent meanings in texts into maps. To the best of our knowledge, the only study to consider different layers of meaning was conducted by Mohr and colleagues (2013), in which they analyzed the meaning hidden in U.S. National Security Strategy reports from 1990 to 2010. Mohr and colleagues used topic modeling to identify different scenic situations in their text corpus, and they used these situations to compare the rhetoric of different U.S. administrations—that is, the co-occurrence of nouns (actors) and verbs (actions). Their analysis was inductive and more interpretive regarding how U.S. administrations legitimated their policies. However, they did not consider the entire text corpus and the actual communication structure, specifically the grammatical connection of subjects, verbs, and objects in semantic triplets, which would have uncovered manifest narratives of social action in and across time. In addition, Mohr and colleagues used the scenic situations as filters with which to construct different co-occurrence maps. This implies that they did not consider the potential dynamic interlacement of different meaning layers. We want to illustrate this dynamic interlacement, and we believe our approach is a valuable contribution to the contemporary trend of using sophisticated text-mining tools.

In the following section, we conduct a demonstration study about the development of the macro-framing of corporate responsibility in the United States from 1950 to 2013. In this context, we apply a reliable, accurate, and free downloadable software application that enables



simultaneous consideration of two layers of textual meaning within maps: the interlacement of communication structures (i.e., operationalized with grammatical parsing) and their semantic surroundings (i.e., operationalized with topic modeling).

### **3. DEMONSTRATION EXAMPLE: BUILDING MAPS FROM COMMUNICATION STRUCTURES AND THEIR SEMANTIC SURROUNDINGS—CORPORATE RESPONSIBILITY IN THE UNITED STATES**

#### *3.1. Theoretical Background*

In our demonstration example, we investigate corporate responsibility as an ongoing discussion in modern society (Meyer, Pope, and Isaacson 2015). One stream of research studying corporate responsibility takes an institutional perspective (Brammer, Jackson, and Matten 2012) and suggests that regulations (i.e., laws and bylaws) in national contexts are closely intertwined with the enactment of corporate policies. This implies that the configuration of regulations evokes a specific meaning regarding corporations' responsibilities (Kang and Moon 2012; Kaplan 2015). Consequently, studies have tied changes of responsibility to changes in the regulatory context (Hiß 2009; Kang and Moon 2012; Matten and Moon 2008) and proposed that the meaning of responsibility shifts in correspondence with the regulatory context.

However, to understand corporate responsibility as an institution in modern society, examining regulations is rarely sufficient because regulations represent a historical process (Brammer et al. 2012; Campbell 2007) and potentially mask the development of meaning. Institutional theory has highlighted the role of language in depicting changes in the meaning of institutions (e.g., Berger and Luckmann 1967; Cornelissen et al. 2015; Phillips, Lawrence, and Hardy 2004). To trace the meaning of corporate responsibility over time, we focus on the macro-framing of corporate responsibility in the context of the United States. We chose this national context because the responsibility of corporations has been discussed in the United States for a long time (Kaplan 2015), and the existing literature enables us to derive an easy and comprehensible research focus.

The literature assumes that the responsibility of corporations in the United States has not significantly shifted over time. Regulations ensure

corporate responsibility with a strong focus on instrumental needs and shareholders' claims (Maignan and Ralston 2002). This implies that corporations' additional social engagement is considered desirable but aligns with a strict individual cost-benefit analysis (McWilliams and Siegel 2001). According to the existing literature, only a moderate shift might have occurred in the 1990s (Kang and Moon 2012; Marens 2012; Shamir 2011). In line with neoliberal ideologies, the focus on shareholder value management was strengthened in a regulatory manner (Amable 2011; Davis 2009). The global financial crisis in 2007 and 2008 in turn led to a moderate reconsideration of shareholder value-driven corporate governance. However, this reconsideration did not result in a path-shifting change but in regulative reforms that aimed to ensure the smooth functioning of markets (Davis 2009; Dobbin and Jung 2010). The financial crisis heightened public focus on corporate responsibilities, which were now not only considered desirable but an integral part of a corporation's strategy (Kang and Moon 2012).

In our demonstration example, we explore the meaning of corporate responsibility in public discourse over time and study whether this development corresponds with the existing literature's assumptions. We investigate meaning by building on the idea of macro-framing in public discourse. According to Cornelissen and Werner (2014), linguistic framing first can consider communication structures, that is, semantic triplets, comprising actors (Who should take responsibility?), actions (What kind of actions are associated with responsibility?), and objects (What kind of responsibility should be assumed?). Second, the meaning of communication structures is influenced by thematic orientations in discourse (i.e., semantic surroundings in our terminology). Together, both layers of meaning enable an inductive but still formal in-depth analysis that facilitates first interpretations and, by remaining open for subsequent close reading, provides a beneficial road map for a deeper consideration of responsibility's multiplexity (cf. Abend 2014). Consequently, we consider our demonstration example to be well suited to illustrate the potentials of our methodological framework.

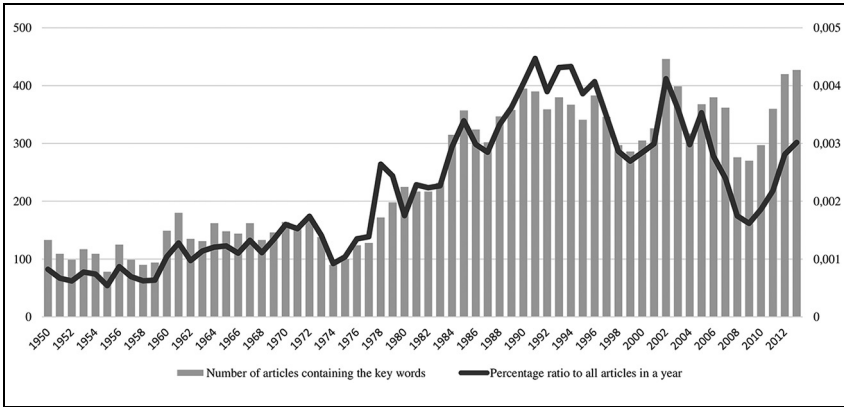
### 3.2. *Text Corpus*

To portray the development of the framing of corporate responsibility in the United States over time, we chose the narrative content of two nationwide newspapers distributed daily. Accordingly, we collected

newspaper articles published by the *New York Times* and the *Washington Post* between 1950 and 2013.<sup>2</sup> We collected our text corpus using the ProQuest database for historical newspapers (<http://www.proquest.com>) for 1950 to 1977 and the Nexis database (<https://www.nexis.com>) for 1978 to 2013. With corporate responsibility, we focus on the first-order dimension that Abend (2014) considered the communicatively manifest dimension of morality. These moral obligations are predominantly reflected in utilization of the term *responsibility* (de George 1999). In contrast to this publicly available dimension, Abend (2014) considered the moral background—namely, the metaphysical premises that buttress morality (e.g., whether morality is buttressed by a secular or sacral fundament)—to be the second-order dimension. In our demonstration example, we focus on communicated responsibilities; to do so, we searched the two newspaper databases for all available articles from the business sections that contained the key words *responsibility* and/or *responsibilities*. The key words are connected to corporations when an article linguistically expresses that a corporation or its representatives are considered responsible. To extract the articles that fulfilled this criterion, we followed recent developments in social science and used a naïve Bayes classifier (cf. Nardulli, Althaus, and Hayes 2015). The classifier was trained with 1,000 manually classified newspaper articles. Tests with 1,000 unclassified articles yielded a classification model that correctly detected 97.3 percent to be relevant. Overall, our final text corpus encompasses 15,371 articles (for descriptive statistics, see Figure 1). Because only the *New York Times* makes available the total number of articles they publish each year, we calculate a relative measure only using this newspaper.

### 3.3. *Communication Structure: Grammatical Parsing and Named Entity Recognition*

Our first goal is to operationalize the communication structure in sentences. In line with our research endeavor, we operationalize the communication structure as a grammatical connection of subjects, verbs, and objects. In the context of narrative text genres, the most common way to do so is to draw on the idea of semantic triplets (Franzosi 1989; Roberts 1989). As this section progresses, we will describe how grammatical parsers extract semantic triplets. Before a grammatical parser can work, the text must be split into sentences and the sentences into words. These

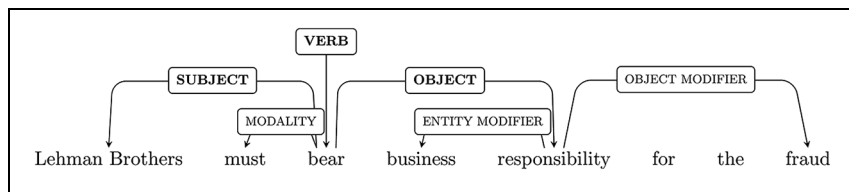


**Figure 1.** Articles referring to corporate responsibility.

words in turn must be prepared using part-of-speech tagging. A part-of-speech tagger classifies words into parts of speech, such as nouns, verbs, adverbs, and prepositions (Brill 2000; Voutilainen 2004). Part-of-speech tagging for English also includes linguistic attributes (e.g., nouns in plural, adverbs in superlative, and verbs used in the third-person singular in present tense; Santorini 1990; Trost 2004).

After this preparation step, researchers can consider two types of grammatical parsing. First, parsers can be built on *phrase structure grammar* (e.g., Matthews 1981). Such constituency parsing highlights that sentences consist of phrases encompassing one or more words. Second, parsers can be built on a *dependency grammar* (e.g., Mel'čuk 1988; Nederhof and Satta 2013). Dependency parsing highlights the grammatical connection of words in sentences and considers grammar to be a network of dependencies between the words in a sentence (Carroll 2004). One word represents the root of a sentence, and all other words depend on this root directly or transitively via other words. Dependency parsers annotate dependencies with the respective grammatical function (Chen and Manning 2014; de Marneffe, MacCartney, and Manning 2006) and are well suited to operationalize the connection of subjects, verbs, and objects in semantic triplets.

Figure 2 depicts a subject–verb–object triplet. Dependency parsing reveals that the subject *Lehman Brothers* represents the actor, the verb *bear* describes the action, and *responsibility* is the object. The units arranged around this triplet serve as modifiers and provide more detailed



**Figure 2.** Grammatical coding of an example sentence.

information. For example, the *entity modifier* for *responsibility* is *business* because it specifies the object. The object *responsibility* is further specified by the *object modifier*, *fraud*. The *temporal orientation* of the verb *bear* is present tense. The modal auxiliary verb *must*, which depends on the verb *bear*, provides the *modality* of the sentence. For example, modality can be used to express obligations, intentions, or abilities/possibilities (Popping and Roberts 2015; Roberts et al. 2010; Vicari 2010). Here, obligations (e.g., *must*) refer to morality, and intention (e.g., *will*) expresses the will toward action, whereas ability/possibility (e.g., *can*) only codes the conditions that enable actions (Vicari 2010).

Note that in our demonstration example, the keywords *responsibility* and *responsibilities* are invariably the objects of the semantic triplets. To comprehensively analyze the framing of corporate responsibility, instead of just considering the single object *responsibility*, we focus on the entity modifiers and object modifiers of this object. This means the triplets we will present have the following structure: “subject–verb–entity modifiers of responsibility” and “subject–verb–object modifiers of responsibility.” Finally, the grammatical position of subjects is regularly occupied with named entities (e.g., Paul or Goldman Sachs). *Named entity recognition* software supports the automatic assignment of grammatical subjects to more meaningful conglomerates, such as *person* or *organization* (Carstensen et al. 2010; Florian et al. 2003; Sutton and McCallum 2006).

The underlying technique enabling the functioning of a sophisticated natural language processing tool, such as part-of-speech tagging, grammatical parsing, and named entity recognition, is supervised machine learning. Machine learning focuses on computer-supported processes, which build up real-world structures based on experience with data (Langley 1996; Mitchell 1997). In terms of part-of-speech tagging, grammatical parsing, and named entity recognition, this implies usage

of “a large and principled collection of natural texts” (Biber, Conrad, and Reppen 1998:4). The term *text* encompasses originally written documents (e.g., newspaper articles). A collection of texts is *principled* if the corpus adequately represents a specific use of language (e.g., professionally edited language of different text genres).

To train text-mining tools using supervised machine learning, these corpora are transformed into *treebanks*, that is, they are manually or semi-manually annotated with linguistic features (e.g., parts of speech, named entity, and dependency labels). The software learns from the sample sentences by considering the values assigned to the features. This means the software searches for how the features occur within various textual contexts to identify the textual structures that might, for example, predict the grammatical function of words or the occurrence of named entities in sentences. In this way, training means the algorithm processes the examples considering the features. The result of such training is called a *model*. The model can then be applied to unknown sentences to predict parts-of-speech tags, grammatical dependencies, or named entities (i.e., classification). The accuracy of parts-of-speech taggers, dependency parsers, and named entity recognizers thus depends on the treebank used. For example, software that was trained using professionally edited texts (e.g., newspaper articles) will likely perform well on texts of similar genres but may be weak on texts possessing flawed grammar or high linguistic ambiguity (e.g., Twitter posts). Therefore, researchers must use a model that fits the text genre being analyzed.

To extract the semantic triplets used in our demonstration example, we used the sophisticated Stanford CoreNLP (Manning et al. 2014).<sup>3</sup> We do so because the Stanford part-of-speech tagger comes with a model trained and tested on the Penn Treebank. The Penn Treebank is a corpus consisting of newspaper articles from the *Wall Street Journal*, so we believe the part-of-speech tagger model is appropriate for analyzing our newspaper articles. In tests on the Penn Treebank, it achieved an accuracy of 97.24 percent (Toutanova, Klein, and Manning 2003). Furthermore, the Stanford parser comes with a model trained and tested on the Penn Treebank. This test shows that the parser managed to parse sentences with an accuracy of 92.20 percent (Chen and Manning 2014). To classify whether the subject of each semantic triplet is a person or organization, we used the Stanford named entity recognizer. The model of the named entity recognizer was trained on the MUC 6 and MUC 7 treebanks, which also consist of newspaper articles from the *Wall Street*

*Journal*. Subsequently, the model was tested on Reuters newswire articles and achieved an accuracy of 86.86 percent (Finkel, Grenager, and Manning 2005). Therefore, we are convinced the model of the named entity recognizer fits with the text corpus used in our demonstration example.

Nevertheless, the actual accuracy of part-of-speech tagging, dependency parsing, and named entity recognition depends on the concrete text corpus analyzed. Consequently, in the following, we validate our software application's results by comparing them to those of trained coders.

### 3.4. *Semantic Surrounding: Topic Modeling*

Our second goal is to operationalize the broader semantic surroundings of semantic triplets (Kahl and Grodal 2016; Khaire and Wadhwani 2010; Kirchner and Mohr 2010; Wagner-Pacifici et al. 2015). Communication structures do not carry one static meaning; rather, their meaning is dynamic and depends on the semantic surroundings they are embedded in and vice versa (Krippendorff 2004; Popping 2012; Vossen 2004). Topic modeling, which builds on the idea of unsupervised machine learning (Langley 1996; Mitchell 1997), is an increasingly common approach to discover hidden semantic patterns in large text corpora (Mohr and Bogdanov 2013). Comparable to part-of-speech tagging, dependency parsing, and named entity recognition, this kind of machine learning aims to build real-world structures from text data. However, in contrast to these tools, topic modeling is a class of generative, Bayesian probabilistic models that does not require annotated tree-banks (Evans and Aceves 2016). Instead, the most common algorithm for topic modeling, the latent Dirichlet allocation algorithm (DiMaggio et al. 2013; Fligstein, Brundage, and Schultz 2017; Mohr et al. 2013), inductively discovers hidden semantic patterns in unannotated texts by detecting significant structures in word co-occurrences (Blei 2012; Blei et al. 2003).

To extract the semantic surroundings of the semantic triplets in our demonstration example, we used the “lda” Python package,<sup>4</sup> which implements the latent Dirichlet allocation algorithm (Blei et al. 2003). We define the semantic surrounding of triplets as being constrained by the paragraph in which the key words *responsibility* or *responsibilities* occur. To ensure the semantic surrounding is not mixed with the

semantic triplet elements, we excluded all sentences within the paragraphs that contain the key words *responsibility* or *responsibilities*. Furthermore, to achieve interpretable semantic patterns, our analysis includes only adjectives and nouns (excluding proper nouns). Therefore, instead of using stop words, we used the Stanford part-of-speech tagger (Manning et al. 2014; Toutanova et al. 2003) to exclude all word classes that do not contribute content meaning to the semantic patterns (Voutilainen 2004).

Application of the latent Dirichlet allocation algorithm requires pre-definition of the number of semantic patterns that should be detected (Blei 2012). This practice is appropriate if the defined number of semantic patterns yields interpretable and analytically useful results (e.g., DiMaggio et al. 2013). To establish transparency regarding our topic modeling procedure, we present further details and descriptive statistics about our model’s fit in the following.

## 4. RESULTS

### 4.1. *Assessing Reliability and Accuracy*

4.1.1. *Communication Structure*. To function properly, it is critical that our software application accurately detect grammatically connected subjects, verbs, and objects as well as entity modifiers and object modifiers of responsibility (*extraction of semantic triplets*). Second, it is critical that named entities are correctly labeled as persons and organizations (*classification of named entities*). The software’s output is suitable for conducting tests of reliability and accuracy (for more details on how to calculate reliability and accuracy measures, see the software’s manual).

In this section, we discuss how reliable and accurate the software was in semantic triplet extraction and named entity classification. The gold standard for these two assessments, which we used to validate the quality of our software, was provided by trained coders. To validate both tasks, we randomly singled out 1,000 sentences (one semantic triplet per sentence) that included our key words and asked each coder to verify the results of our software application. To secure the manual coding procedure’s reliability and accuracy, the coders mutually verified their results and corrected deviations after unanimous agreement.

Table 1 reports results of our reliability and accuracy assessment. The precision score for semantic triplet extraction displays what



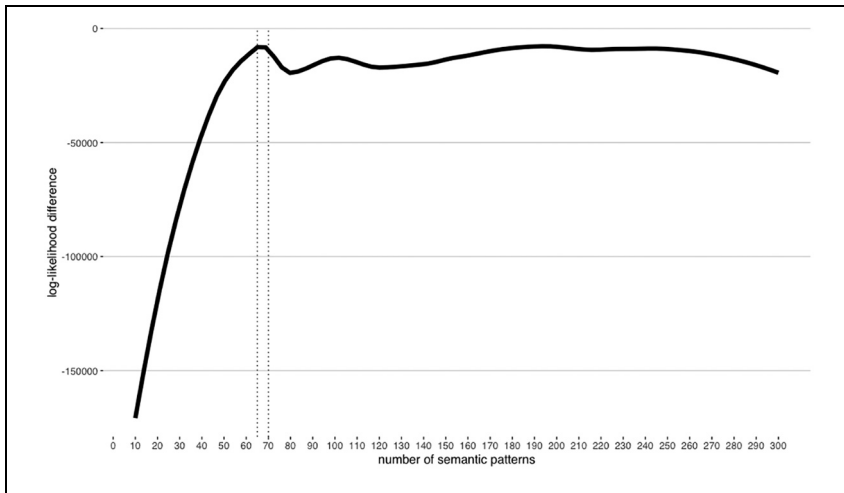
**Table 1.** Reliability and Accuracy of the Semantic Triplet Extraction and Named Entity Classification

Task	Precision	Recall	<i>F</i> -measure
Semantic triplet extraction	96.44	97.36	96.90
Named entity classification	98.42	98.79	98.60

percentage of the semantic triplet elements detected by our software application belongs to a semantic triplet (correctness test). The recall score for semantic triplet extraction indicates what percentage of all semantic triplet elements that appear in the sentences was actually detected by our software application (completeness test). The precision score for named entity classification displays what percentage of the named entities detected by our software application is persons and organizations (correctness test). The recall score for named entity classification indicates what percentage of all persons and organizations, as subjects, that appear in the texts was actually detected by our software application (completeness test). We combined both measures by reporting the commonly used *F*-measure (Manning and Schütze 2000), which depicts the overall performance of both tasks.

**4.1.2. Topic Modeling.** Applying the latent Dirichlet allocation algorithm requires predefinition of the number of semantic patterns that should be detected (Blei 2012), so it is critical to make this step as transparent as possible. Hence, we first tested which number of semantic patterns technically appears as optimal. To this end, we used a cross-validation based on the log-likelihood, which is regularly used to evaluate topic modeling (Griffiths and Steyvers 2004). With this validation, we determined the number of semantic patterns at which the increase in model improvement stabilizes (see Figure 3). Our software includes the option to conduct a cross-validation using log-likelihood (for more details, see the software’s manual).

In our case, the log-likelihood approach revealed that a configuration between 65 and 70 semantic patterns is technically optimal. We thus followed recent suggestions and computed topic models with semantic patterns ranging between 65 and 70 to examine which model produces semantic patterns that are interpretable and analytically useful (see



**Figure 3.** Cross-validation of different topic models.

DiMaggio et al. 2013; Nelson 2017). In the end, we decided on a model with 70 semantic patterns.

Second, we followed recent suggestions to build on observations or patterns that fit the focus of the analysis (Breiger et al. 2018; DiMaggio et al. 2013; Nelson 2017). Because we focus on corporate responsibility as a societal phenomenon, we chose semantic patterns related to societal issues. To this end, we again followed recent advice from social scientists (Breiger et al. 2018; Mohr and Bogdanov 2013) and returned to our textual material to qualitatively confirm the plausibility of patterns and inductively give patterns semantic labels (Nelson 2017). Afterward, we excluded semantic patterns referring to macro-economic aspects, industries, and sheer functional processes and accordingly reduced the number of patterns to 32 (see Table 2; for the 38 unused patterns, see the Appendix).

Third, we pooled patterns that depict similar meanings and ended up with six semantic groups: management, standards and performance, shareholder issues, employment, social issues, and misconduct and risk. In our demonstration example, we use these semantic groups as the relevant semantic surroundings of triplets.

We paralleled a qualitative confirmation of semantic groups with a statistical step that ensured the groups encompassed semantic patterns

**Table 2.** Semantic Patterns in Six Groups

Semantic Groups	Semantic Patterns	Top Words
Shareholder issues	Stock market and investors	stock, investors, shares, market, companies, analysts, price, percent, stocks, company, public, research, investment, investor
	Stock market and profits	company, year, percent, last, analysts, earnings, share, sales, years, quarter, profit, revenue, profits, analyst
	Shareholders and acquisition	shareholders, offer, takeover, company, shares, stock, board, management, shareholder, time, percent, merger, share, control
	Shareholders and investors	companies, shareholder, corporate, shareholders, social, company, investors, institutional, issues, group, responsible, proxy, investment, corporations
Management	Investors	capital, private, investment, equity, firm, deal, investors, company, group, partners, buyout, venture, companies, deals
	Board of directors and management	board, directors, committee, director, chairman, members, management, boards, corporate, company, independent, member, governance, former
	Board of directors and management	management, association, company, board, owners, associations, manager, members, community, new, contract, services, member, time
	Executive officers and management	president, company, executive, chairman, vice, chief, officer, corporation, years, yesterday, old, business, committee, post
	Executive officers and management	president, vice, executive, senior, director, chief, division, manager, company, officer, general, new, group, operations
Employment	Executive officers and management	executive, chief, company, chairman, president, officer, management, top, years, yesterday, time, job, executives, year
	Executive officers and management	executives, executive, chief, corporate, company, top, former, chairman, people, business, companies, officers, senior, president
	Managers and career	years, president, university, year, business, job, degree, old, school, career, last, chairman, company, vice
	Organization of work (general)	management, managers, people, employees, companies, company, new, top, manager, years, organization, team, corporate, job
	Organization of work (subunits)	office, management, department, staff, personnel, offices, branch, operations, ball, manager, system, money, general, cash

*(continued)*

**Table 2.** (continued)

Semantic Groups	Semantic Patterns	Top Words
Social issues	Social engagement	foundation, year, people, groups, organizations, nonprofit, organization, group, philanthropy, foundations, public, charity, charitable, last
	Social engagement	business, corporate, companies, corporations, social, company, corporation, employees, way, community, mission, profits, programs, new
	Protection of minors	tobacco, industry, campaign, smoking, kids, ads, companies, advertising, children, cigarette, beer, marketing, public, percent
	Protection of labor	safety, workers, company, waste, factory, factories, labor, suppliers, companies, environmental, conditions, year, employees, violations
	Working time	job, time, people, good, work, way, jobs, new, day, years, sure, months, thanks, things
	Work-life balance	family, home, children, time, father, wife, son, day, young, years, life, parents, child, year
	Employees	women, employees, work, percent, companies, men, company, job, employee, employers, woman, jobs, workers, workplace
	Unions	workers, union, labor, jobs, employees, work, unions, job, worker, contract, wage, wages, strike, benefits
	Business ethics	business, ethics, professor, corporate, ethical, company, people, university, public, code, American, problem, good, companies
	Standards and performance	study, time, problem, problems, standard, fact, number, example, performance, management, important, standards, result, quality

(continued)

**Table 2.** (continued)

Semantic Groups	Semantic Patterns	Top Words
Misconduct and risk	Risk, prices, and financial crisis	bank, financial, risk, prince, executive, chief, derivatives, role, former, executives, markets, treasury, crisis, management
	Risk, losses, and financial crisis	bank, chief, regulators, risk, trading, global, financial, executives, diamond, office, British, investment, losses, former
	Losses, industry, and financial crisis	problems, last, losses, year, problem, week, bad, loss, days, industry, analysts, money, big, financial
	Subprime crisis	mortgage, loans, loan, securities, debt, financial, investors, credit, mortgages, market, fannie, mae, housing, rate
	Bankruptcy and overextension	bankruptcy, company, plan, creditors, court, protection, chapter, assets, debt, filing, agreement, agency, trustee, claims
Legal investigation and lawsuit	Legal investigation and lawsuit	company, investigation, former, settlement, executives, general, yesterday, commission, complaint, securities, officials, suit, regulators, practices
	Legal investigation and lawsuit	court, case, law, lawyers, legal, judge, suit, lawyer, cases, lawsuit, decision, lawsuits, federal, liability
	Legal investigation and lawsuit	case, government, criminal, fraud, federal, charges, guilty, prosecutors, justice, department, former, investigation, scheme, court

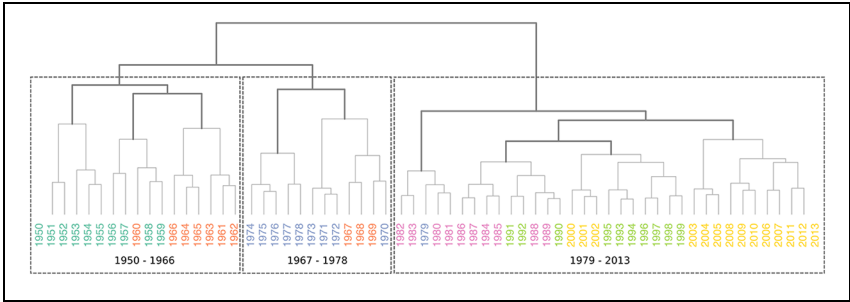
**Table 3.** Topic Modeling: Ingroup Versus Outgroup Comparison (Pearson's Correlation Coefficient)

Semantic Groups	Ingroup Correlation	Outgroup Correlation
Pattern distribution		
Shareholder issues	.23	.13
Management	.17	-.12
Employment	.14	.03
Social issues	.46	.13
Standards and performance	—	-.18
Misconduct and risk	.61	.14
Word overlap		
Shareholder issues	.17	.09
Management	.30	.11
Employment	.24	.12
Social issues	.12	.09
Standards and performance	—	—
Misconduct and risk	.21	.07

with similar meanings and trends. We verified whether the occurrence of the semantic patterns pooled into the same group was similar over time (pattern distribution) and that the semantic patterns were similar regarding the top 50 words, which are most representative of each pattern (word overlap). To accomplish this, we checked whether the average ingroup correlations of the semantic patterns was higher than the average outgroup correlations. The outgroup is defined as all other semantic patterns. The results revealed that the ingroup correlations were higher than the outgroup correlations (see Table 3). Semantic patterns pooled into a group tend to occur simultaneously over time and share a considerable amount of words. However, high ingroup correlations should not be expected. Even if patterns pooled in a semantic group tend to be similar in pattern distribution and word overlap, social scientists should bear in mind that patterns in semantic groups differ significantly enough to represent the facets of an overarching phenomenon.

#### *4.2. Corporate Responsibility in the United States: Maps from Communication Structures and Their Semantic Surroundings*

In this section, we analyze and visualize the macro-framing of corporate responsibility by using several inductively generated maps to trace its development. We created a graphic representation to visualize the

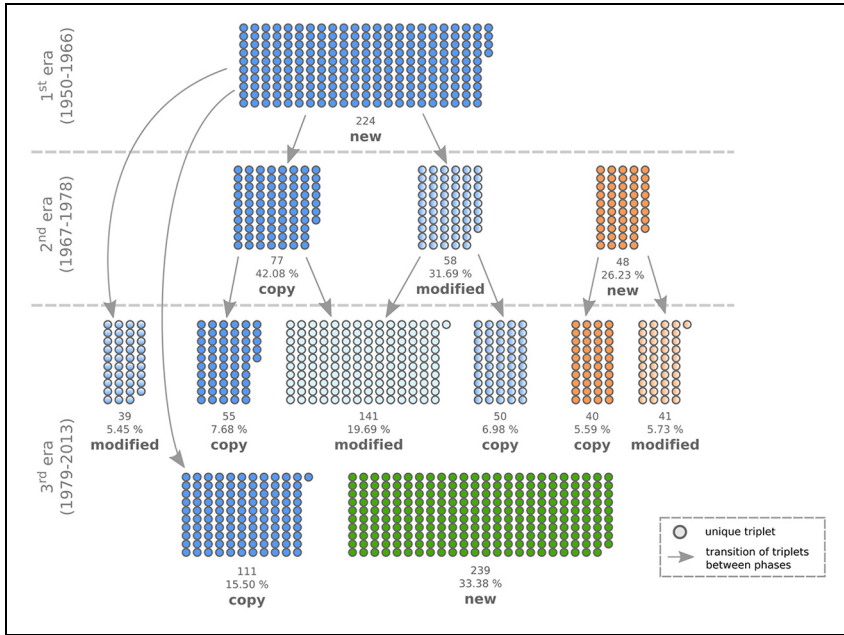


**Figure 4.** Hierarchical clustering.

stability and change of communication structures by identifying eras of similar semantic triplet usage. We use this graphic representation, which focuses on the count of semantic triplets as a relevant textual characteristic, to guide our subsequent creation of maps. In detail, we drew on a hierarchical clustering approach using the relative distance of the semantic triplets' usage between the years 1950 and 2013 (see Figure 4). Years that show short distances to each other were shaped by similar usages of semantic triplets. We calculated the similarity of years as their distance in a multidimensional vector space, in which each semantic triplet represents one dimension (e.g., person-fulfill-shareholder; organization-have-social). All dimensional values were normalized by the absolute count of given triplets. To smooth the outliers in our time series, we used the moving average ( $k = 3$ ). We then calculated the Euclidian distance between the years. To detect clusters, we used the ward.D hierarchical clustering algorithm to reduce the dimensionality of our multidimensional vector space (see Manning and Schütze 2000; Mohr 1998).

Our findings reveal that macro-framing of corporate responsibility in the United States went through three basic eras: early era one, encompassing 1950 to 1966; middle era two, from 1967 to 1978; and late era three, from 1979 to 2013. Our clustering also reveals that this last era is subdivided into four additional subclusters: 1979 to 1983, 1984 to 1992, 1993 to 2002, and 2003 to 2013.

Our first analysis step provides additional insights regarding corporate responsibility. The macro-framing of responsibility has significantly changed over time, but the 1990s and 2000s do not seem to be outstanding in this regard. This contrasts with existing literature that claims the 1990s and 2000s marked a moderate shift in the development



**Figure 5.** Map 1: Transition of semantic triplets across eras.

of corporate responsibility (cf. Kang and Moon 2012; Marens 2012; Shamir 2011). Our analysis shows that the 1990s and 2000s are just a subcluster within an era that stretches from 1979 to 2013. These results point to the relevance of thinking about additional events and developments that might explain the significant shifts we encounter in the framing of corporate responsibility. For example, one might consider the effects that the first economic slump in the world economy after World War II in 1967 to 1968 (Malsberger and Marshall 2009) or the beginning of economic liberalization in the 1980s (Simmons, Dobbin, and Garrett 2006) might have had on the responsibilities ascribed to corporations.

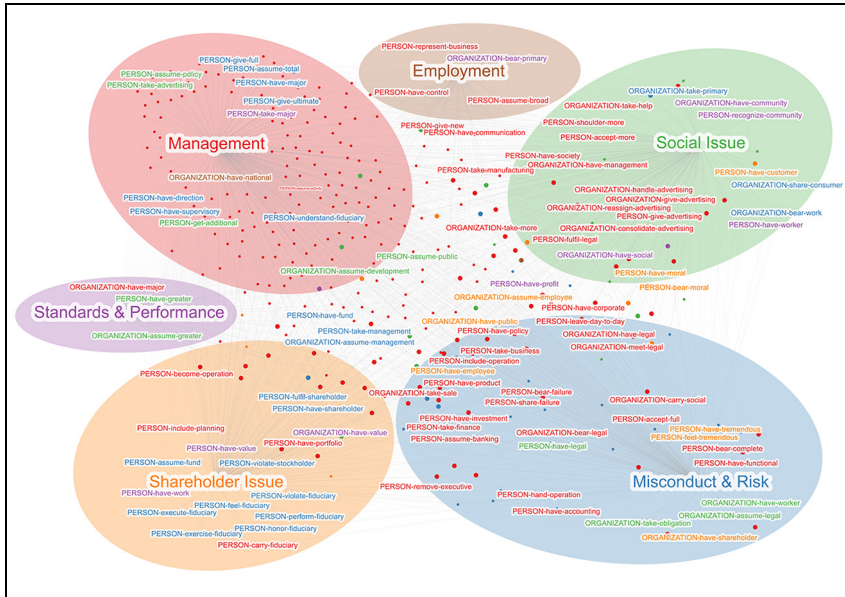
In the following, we use the clustering results to deepen our analysis by creating several maps. First, we take the three main clusters and investigate how the semantic triplets developed over time. To achieve this, we focus on the number of unique semantic triplets per main era as a relevant textual characteristic and then map the connectedness and proportion of semantic triplets in time. Taking a bird's-eye view, this map visualizes when and how the semantic triplets emerged in and moved through time (see Figure 5).



In Figure 5, the semantic triplets on the top in blue are those that existed in the era from 1950 to 1966. The arrows depict the transition of these triplets into other eras. Triplets from era one could be copied to era two, meaning the triplet remained completely the same. However, triplets from era one could also be modified in their transition to era two, meaning the object remained constant, but the subject or verb changed. In addition, several new semantic triplets appeared for the first time in era two (red). Note that multiple semantic triplets from era one did not exist in era two but entered era three as copies, or at least in a modified form. Era three also contains triplets transmitted as copies from era one to era two to era three, as well as triplets that are copies of triplets that were modified during their transmission from era one to era two. Era 3 also has triplets that are copies or modifications of triplets that first occurred in era two. Finally, some completely new semantic triplets (green) emerged in era three.

Our first map demonstrates that the changes in macro-framing of corporate responsibility did not happen gradually, as the existing literature suggests (cf. Kang and Moon 2012). Instead, the map shows that only 7.68 percent of all triplets in era three are direct copies of triplets that also occurred in eras one and two. At least 20.95 percent of triplets in era three did not occur in era two but were transmitted from era one. Era three also encompasses triplets that are copies (5.59 percent) or modifications (5.73 percent) of semantic triplets that first occurred in era two. Additionally, some triplets in era three are modifications of triplets that were previously copied to era two (19.69 percent) or changed during their transmission (6.98 percent). Finally, 33.38 percent of triplets in era three are completely new. In summary, this map reveals that the discourse around corporate responsibility significantly expanded over time (Meyer et al. 2015).

We now turn to era three, which contains many semantic triplets from eras one and two, to study the development of semantic triplets that were copied or modified over time in more detail. To this end, the next map focuses on the strength of association of semantic triplets with the six semantic groups (see Figure 6). This map visualizes proximity, connectedness, and scale of textual characteristics in the text corpus, namely, the nature and strength of interlacement of the semantic triplets and their semantic surroundings. We use a network analysis (Wasserman and Faust 1999) to map this association in era three. The weight of the edges stems from the degree of association a semantic triplet has with the six



**Figure 6.** Map 2: Semantic associations of triplets in era three (large nodes and nodes with labels changed their semantic association over time).

semantic surroundings. The larger nodes (including the nodes marked with a label) indicate triplets that changed their semantic association over time. In other words, for each triplet, we mathematically checked whether the relative semantic association with the semantic surrounding changed compared to previous eras. The small nodes did not change their semantic association over time. The color of the nodes indicates their semantic association within an earlier era. To keep our map as clear as possible, we only show labels for triplets occurring most frequently in discourse. For visualization, we use the Force Atlas layout, which supports the drawing of easily readable graphs (Kobourov 2012). To make our map more interpretable, we will provide some descriptive statistics and quotes from newspaper articles.

Map 2 (Figure 6) shows a considerable and previously rarely considered degree of correspondence between events in the economic history of the United States and the meaning of corporate responsibility. Economic liberalization, which started in the 1980s, put corporations in a complex situation: They were considered freely acting individual units on markets that should balance shareholder interests and the stakes of

society and the environment (Kaplan 2015). From the 2000s onward, however, corporations were criticized for their shareholder orientation, which was cited as the cause of serious misconduct and distortions in the economy and society as a whole (Streeck 2014). In contrast, the booming economy from 1950 to 1967 and the reconstruction of a peacetime production regime led corporations to focus on managing their business while the national state ensured social welfare (Boyer and Hollingsworth 1998; Vidal 2011). The first economic slump in 1967 to 1968 caused a slight reconfiguration of these tendencies. The following years were characterized by a search for more flexible production and efficiency as well as a focus on human resources (Vidal 2013) that could improve competitiveness in the global market (Pedersen 2010).

Map 2 (Figure 6) shows this change in corporations' expected role in society: 46.3 percent of triplets changed their semantic association over time. The strong focus on management in earlier eras had changed by era three: For example, in era three, 49.5 percent of the triplets now associated with misconduct and risk (e.g., person-bear-failure) and 55.9 percent of triplets now associated with social issues (e.g., organization-have-management) were previously associated with management. These results and the following newspaper quotes illustrate the changed focus of corporate responsibility, namely, a transition from mainly managing the corporation to considering the effect of business on society:

While store executives at the week's sessions conceded that many salespeople were badly chosen and badly trained, . . . **[Person] bears** a direct *responsibility*, not only for **failures** in personnel practices, but for other major faults, it was said. (*New York Times*, January 15, 1959)

The directors also voted to release the results of internal investigations into the trading losses, which largely fault other top executives for the problems. . . . "**[Person] bears** ultimate *responsibility* for the **failures** that led to losses," the board said in a statement. . . . Still, the trading losses, which have swelled to more than \$6 billion, have cast a long shadow over the board and management of the [Organization]. (*New York Times*, January 17, 2013)

The [Organization] reported yesterday a . . . increase in sales. . . . This combines with domestic and Canadian operations the overseas subsidiaries for which the **[Organization] has management responsibility**. (*New York Times*, April 13, 1966)

Most such take-back programs are run by local or other government agencies. But increasingly there are calls to make the pharmaceutical industry

pay. “We feel the **[Organization]** that profits from the sales of these products should **have** the financial *responsibility* for proper **management** and disposal,” said [Person]. (*New York Times*, December 7, 2012)

In a similar vein, in era three, 50 percent of triplets previously associated with standards and performance were related to social issues (e.g., organization-have-social), and 25 percent were associated with shareholder issues (e.g., person-have-value). In addition, 28.6 percent of the copied or modified triplets associated with shareholder issues in era three were previously associated with misconduct and risk (e.g., person-violate-fiduciary).

Federal judge in Wilmington, Del., appointed a receiver yesterday to take control of the [Organization], . . . The suit charged that **[Persons]**, all officers or directors of [Organization], had **failed** to live up to their **fiduciary responsibilities** in running the company. [Person], who controls at least a dozen corporations and banks through [Organization], is under indictment in New York on charges that he lied about certain transactions. (*New York Times*, July 13, 1968)

Separately, in the developing acquisition battle, the [Organization A] filed a suit in United States District Court in New York yesterday seeking more than \$25 million in damages from [Person of Organization B], and its directors. The suit charged that the retailer’s **[Persons]** had “publicly blackened” the [Organization A] name and had **violated fiduciary responsibility** to shareholders by seeking to block the [Organization A] without shareholder approval. The suit asked for a court order enjoining [Organization B] from trying to block [Organization A] from making further stock purchases. (*New York Times*, March 2, 1982)

Finally, it is surprising that the semantic surrounding shareholder issues in era three is almost entirely associated with triplets that previously had other associations and triplets that were associated with shareholder issues in earlier eras changed their association dramatically. For example, in earlier eras, triplets concerning employees (e.g., person-have-employee), the public (e.g., organizations-have-public), and moral issues (e.g., person-have-moral) were discussed in the context of shareholder issues. In era three, these triplets tended to be associated with misconduct and risk and social issues.

“There is no question but that an understanding of a company’s expected future economic performance is essential to an accurate evaluation of the current worth of its stock which now is limited to a relatively few professionals,” [Person] said. “A **[Organization]** **has** a *responsibility* to inform the investigative **public** about its own earnings projections and, of

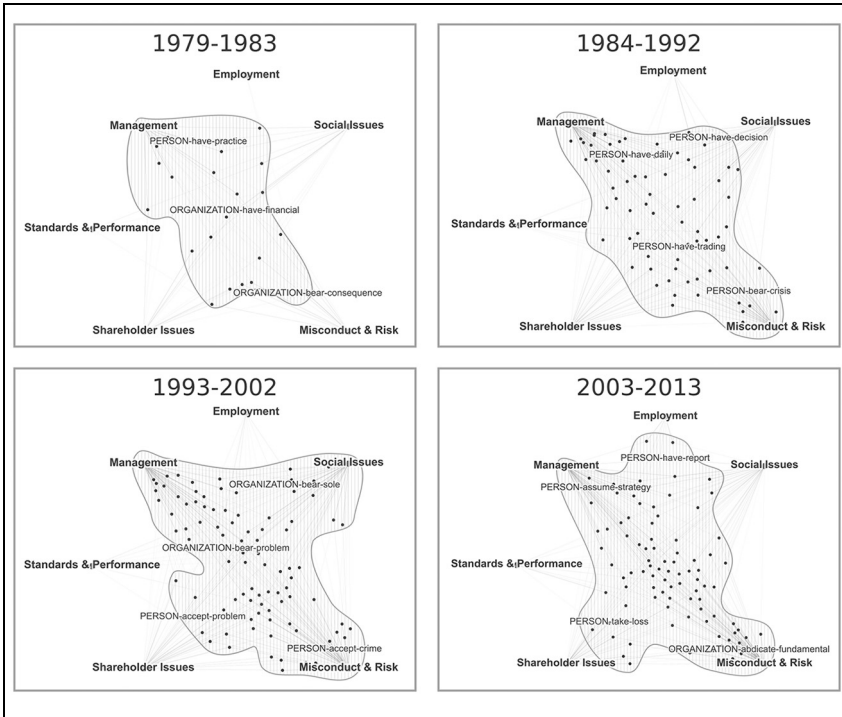
course, that means on a per-share basis. It also has a responsibility to disclose the basis of its earnings forecasts or the various elements entering into it.” (*New York Times*, November 22, 1972)

“The public safety net federal insurance reflects a consensus that banking functions are essential to a healthy economy,” [Person] wrote. “Its presence also implies that **[Organizations]** have unique **public responsibilities**.” As such, he continued, they should remain separate from other types of financial institutions like securities firms and insurance companies. “Subsidiary banking activities should not entail excessive risk of loss and should not impair the impartiality of the credit decision making process.” (*Washington Post*, April 5, 1983)

[Person] explained that the original purpose of establishing a separate company was to remove one step further the liability which the bank might face in any legal action involving the safe deposit business. The **[Organization]**, however, recognizes that as sole owner of the deposit company—it would **have** “at least a **moral responsibility**, and we are unwilling to duck that issue,” [Person] said. During the general discussion it was brought out that the [Organization] is not subject to excess profits taxes in 1953. (*New York Times*, January 28, 1954)

A **[Person]** has **moral responsibility**. Leadership is not a set of skills; it’s a moral issue about doing the right thing. . . . The moral issue is a great one and I think that’s what differentiates smaller businesses from larger ones, because this responsibility is something that has to be taken, it’s not given. You have to feel that you have a contribution to make. It’s the same as being in politics. (*New York Times*, May 20, 2013)

Overall, and in contrast to existing literature on corporate responsibility, we show that era three focused not only on shareholder value-driven corporate governance (cf. Davis 2009; McWilliams and Siegel 2001) but also on issues concerning the social sphere. In addition, our map reveals that communication, in the form of triplets, dramatically changed in the context of shareholder issues. The focus on misconduct and risk is in line with existing literature that describes the financial crisis leading to a moderate focus on corporate responsibility. In our map, however, this focus seems to be strong rather than moderate. Descriptive statistics further uncover a heretofore unrecognized pattern regarding the subjects’ ascribed responsibility: 42.7 percent of all triplets that have a person as the subject are associated with management, whereas 50 percent of all triplets that have an organization as the subject are associated with social issues and misconduct and risk. Given that persons were the



**Figure 7.** Map 3: Expansion of new triplets referring to corporate responsibility in era three.

subject of 67.3 percent of all triplets, these results are noteworthy. Persons are considered responsible for management, whereas organizations are considered responsible for societal issues.

We now turn to semantic triplets that first appeared in era three. Our analysis has revealed that the responsibilities ascribed to corporations significantly expanded during era three. In the following, we trace how corporate responsibility expanded across the four subclusters in era three (see Figure 7). To explore this further, we use the same network analysis technique and map the semantic associations of newly emerging triplets.

Map 3 (Figure 7) shows that between 1979 and 1983, new triplets tended to be associated with management, standards and performance, and misconduct and risk. Between 1984 and 1992, new triplets tended to be associated with management, misconduct and risk, and social. Between 1993 and 2002, new triplets increasingly emerged with a focus

on shareholder issues. Finally, between 2003 and 2013, fewer triplets emerged in the context of social issues (we see a moderate shift toward employment) and shareholder issues but predominantly emerged around misconduct and risk. In line with existing literature, our findings show that shareholder issues gained importance between 1993 and 2002 (e.g., Kang and Moon 2012; Marens 2012). However, we also found an expansion of the macro-framing responsibility related to social issues and misconduct and risk (Streeck 2014). Furthermore, it is surprising that over time, shareholder issues became described by triplets that previously belonged to other semantic contexts, whereas only a few semantic triplets that appeared in era three for the first time became associated with this semantic surrounding.

In summary, our analysis demonstrates that studying the macro-framing of corporate responsibility is a fruitful addition to the existing literature on corporate responsibility and regulations. Our analysis reveals that macro-framing of corporate responsibility significantly expanded from a strong focus on management issues to broader societal issues. Simultaneously, corporations were increasingly considered actors with potential negative effects on society. Taking the existing literature on corporate responsibility into account, our analysis not only offers a more detailed picture, but it demonstrates that the development of regulations is only loosely coupled with the development of meaning.

## 5. DISCUSSION

### *5.1. Text-Mining Tools and In-Depth Analysis: Considering Multiple Layers of Meaning*

Analysis of big data using text-mining tools may yield valuable theoretical insights for the social sciences (Evans and Aceves 2016; George et al. 2016). Accordingly, social scientists have argued that using text-mining tools is a fruitful way to scale inductively grounded close reading (Edelmann and Mohr 2018; Mohr et al. 2015; Muller et al. 2016; Nelson 2017). In particular, recently introduced methodological frameworks suggest either building text analysis exclusively based on the results of a specific text-mining tool (Jockers and Mimno 2013; Mohr et al. 2013; Rule et al. 2015) or applying text-mining tools combined with close reading. Exclusive utilization of text-mining tools, however, has been criticized because such analyses tend to neglect the inductive

character that in-depth analysis based on close reading can provide (Abramson et al. 2018; Chakrabarti and Frye 2017; Nelson 2017).

In this article, we direct attention to text-mining tools' potential to allow an inductive but still formal in-depth analysis of the dynamics of manifest and latent meanings. We propose a framework that calls for the combined application of different text-mining tools that address different layers of meaning. Several studies have focused on the detection of coarse-grained meaning patterns in texts. For example, we have seen an increased use of topic modeling to analyze latent semantic patterns (e.g., Bail 2014; Bonilla and Grimmer 2013; Fligstein et al. 2017; Mohr et al. 2013; Tsur, Calacci, and Lazer 2015). However, text-mining tools can also provide insights about other layers of meaning. By considering semantic triplets such as subject-object or verb-object relations, for instance, text-mining tools allow researchers to detect communication structures that depict sequences of actions over and across time (e.g., Sudhahar et al. 2013; van Atteveldt et al. 2008). Accordingly, we propose a middle way and argue that the combination of text-mining tools allows social scientists to conduct inductive but still formal in-depth analyses that facilitate first interpretations but remain open for subsequent close reading. The combined analysis of different meaning layers appears to be a fruitful endeavor because meaning in different layers is dynamic and mutually dependent (Kahl and Grodal 2016; Khaire and Wadhwani 2010; Kirchner and Mohr 2010; Wagner-Pacifi et al. 2015). In other words, our methodological framework opens space to use the results provided by different text-mining tools for dialectical interpretations that consider different layers of meaning.

We argue that map analysis (Lee and Martin 2015) is especially helpful in utilizing text-mining tools to increase the formalization of text analysis and simultaneously preserve much of its inductive nature. In this context, our framework enables a novel approach that considers different layers of meaning within one map, namely, communication structures and their semantic surroundings. In this sense, we are moving toward a type of formal text analysis that preserves its inductive and interpretative character (Edelmann and Mohr 2018; Mohr et al. 2013, 2015). Our demonstration example indicates that text-mining tools can be used in multiple ways to carve out new insights about social phenomena. During our research, we decided to use hierarchical clustering to inductively uncover shifts in how corporate responsibility is communicatively framed in texts over time. We then used map analysis to



inductively gain more insight into the nature of the shifts we identified. We also show the dynamic interlacement between semantic triplets and their semantic surroundings. Finally, we used map analysis to uncover how the importance of different semantic surroundings has changed over time. Overall, our empirical endeavor, compared to existing studies, allows for a more detailed perspective of corporate responsibility.

In addition, by operationalizing communication structures as semantic triplets, we showed a way to automate the hand-coding procedure of semantic grammar (Franzosi 1989; Roberts 1989). Our article explains the functioning of grammatical parsing as an engine with which to automatically extract and count semantic triplets while reducing the problems of labor- and time-intensive coder training and manual data coding (Franzosi 2010; Franzosi et al. 2012). Grammatical parsing allows extraction of semantic triplets in a potentially cost-effective and accurate manner. However, semantic grammar approaches tend to focus on the sentence level and therefore neglect the textual embedding of sentences (Mohr et al. 2013; Popping 2012). By using topic modeling to consider the semantic surroundings of semantic triplets, we show one potential way to move beyond the sentence level. Our demonstration example proves the value of this approach and shows that semantic triplets changed their gestalt and their semantic surroundings over time and their meaning changed at least slightly.

Yet Lee and Martin (2015:23) criticize topic modeling in their article on map analysis, so we now discuss why our use of topic modeling is not covered by this critique. Note that Lee and Martin do not reject topic modeling in general. On the contrary, following Lee and Martin, topic modeling is appropriate if, first, the labeling of topics does not rely on untransparent *ex ante* coding. Second, topic modeling is appropriate if subsequent analytic steps do not rely on the labels attached to topics. By relying on statistical measures and not labeling, we apply topic modeling differently. First, we use various statistical steps to increase transparency regarding the validation and further processing of topic modeling results. These statistical steps also support the transparency of label attachment. Second, we use the topic modeling results to determine the statistical association between topics (i.e., semantic surroundings in our terminology) and semantic triplets. The appearance of our maps, which are based on these statistical associations, is independent from the labels attached to semantic surroundings. The labels we use help interpret and visualize the results in a comprehensive manner, but they do not affect the maps' appearance. Finally, dialectic consideration of semantic triplets and their

semantic surroundings allows for the ex post interpretations Lee and Martin (2015) favor. In connecting the semantic surroundings provided by topic modeling with the semantic triplets extracted by grammatical parsing, the results of both text-mining tools can meaningfully refer to each other. On the one hand, the extraction of semantic triplets equips social scientists with another way to interpret the semantic patterns provided by topic modeling (Mohr and Bogdanov 2013). On the other hand, topic modeling considers the textual embeddedness of semantic triplets and makes their meaning more interpretable as a result.

Finally, our article supports social scientists in replicating our framework. We provide a fully developed and reliable software application that implements grammatical parsing and topic modeling and enables the large-scale analysis of communication structures (i.e., semantic triplets) and hidden semantic patterns in texts. In particular, grammatical parsing—which has repeatedly been attributed high relevance (Carley 1996; Cornelissen and Werner 2014; Evans and Aceves 2016)—is rarely used in the social sciences (e.g., Goldenstein et al. 2019; Sudhahar et al. 2013; van Atteveldt et al. 2008). This might be because the application of grammatical parsing still requires advanced knowledge in computer science and software development. In this sense, the fully developed and reliable software application is a secondary but beneficial outcome of our article, and it appears to be a useful addition to the reservoir of sociological methods in terms of text analysis.

## *5.2. Directions for Further Development*

Note that the methodological framework we propose is not limited to being realized with grammatical parsing and topic modeling. Rather, our framework suggests a combination of text-mining tools that addresses different layers of meaning (for an overview, see Evans and Aceves 2016; Nelson 2017). However, even if text-mining tools become increasingly sophisticated, they still cannot automatically extract all the types of information that might interest researchers. For example, as Mohr and colleagues (2013) note, the layer of meanings referring to actors' intentions is still hard to incorporate into an automated and formal text analysis. One new type of text-mining tool that could help us solve this limitation is semantic role labeling, which automatically labels semantic roles in a given sentence (Punyakanok, Roth, and Yih 2008). For example, in a linguistic frame (Fillmore 1985;

Langacker 1986) referring to *communication*, linguistic entities can have the semantic roles *speaker*, *addressee*, and *medium*. Admittedly, such an analysis does not uncover actors' intentions, but it could bring us closer to understanding the kinds of situation and roles in which actors operate. However, because Gildea and Jurafsky (2002) report a rather low accuracy for English semantic-role labeling, we do not yet include this tool in our software application.

In our demonstration example, we used named entity recognition to classify named entities as persons or organizations. However, persons and organizations can also be referred to in an indirect manner by using pronouns, which can only be fully understood if the reader refers to preceding sentences. To address this challenge, anaphora resolution tools are being developed. Anaphora resolution is designed to automatically uncover what entity is being referred to by a given pronoun (Ge, Hale, and Charniak 1998; Mitkov 2004). However, Lee and colleagues (2013) report a rather low accuracy for anaphora resolution tools, which is why we have not yet included anaphora resolution in our software application.

Text-mining tools are still unable to disambiguate (What person or organization is being referred to in a text?) or specify (Which demographic characteristics belong to persons and organizations?) named entities. Therefore, additional information must currently be coded manually. However, emerging potential solutions to this limitation in computer and social science suggest two promising avenues. First, social scientists might use online crowdsourcing tools, which assemble a crowd of workers who can be hired on multiple online platforms, to perform brief tasks in a few minutes (for one example in the social sciences, see McCormick et al. 2017). In the context of the disambiguation and specification of actors, these workers, for instance, could be hired to search the internet for additional information (e.g., demographic facts). Second, computer scientists are currently working on technical solutions to make online sources, such as Wikipedia, available on a large scale (Lehmann et al. 2015). These technical solutions could be used to automatically link the output of text-mining tools with information extracted from large online sources.

Finally, a more technical limitation of natural language processing is that although some languages (e.g., English and German) are supported well by this technique, others (e.g., Polish) are not or are even a blind spot in natural language processing (e.g., Bengali). Despite these limitations, the field of natural language processing will likely hold further interesting applications in the future.

**Appendix.** Unused Semantic Patterns

Semantic Groups	Semantic Patterns	Top Words
IT and information	Internet technology	Internet, web, site, online, com, company, information, mail, sites, companies, service, technology, users, search
	Film and music industry	music, film, movie, studio, entertainment, television, films, pictures, universal, box, year, studios, new, movies
	Computer technology	computer, systems, software, system, computers, information, data, technology, companies, new, machines, security, electronic, company
	Television and broadcast	television, network, cable, time, news, media, networks, turner, tv, new, radio, programming, stations, broadcasting
	Telecommunication	phone, telephone, service, companies, company, communications, long, local, distance, telecommunications, customers, services, wireless, lucent
Utility industry	Energy production	power, energy, nuclear, electric, environmental, utility, company, plants, gas, electricity, utilities, coal, carbon, plant
	Oil and gas production	oil, gas, company, petroleum, energy, companies, gulf, production, spill, barrels, exploration, industry, natural, day

*(continued)*

## Appendix. (continued)

Semantic Groups	Semantic Patterns	Top Words
Finance and economy	Payment transactions	credit, card, security, cards, money, account, customers, checks, accounts, payment, merchants, consumers, consumer, fees
	Stock market and securities	exchange, stock, securities, board, commission, trading, market, big, exchanges, members, excise, public, new, member
	Stock market and brokers	securities, firms, firm, brokerage, brokers, broker, investors, customers, clients, commission, investment, financial, industry, business
	Economy and growth	Japanese, American, economy, economic, percent, companies, government, foreign, growth, today, inflation, market, new, tax
	Banking industry	bank, banks, banking, first, trust, chase, bankers, national, loans, loan, money, accounts, American, assets
	Banking industry	banks, bank, financial, government, federal, money, banking, treasury, institutions, reserve, credit, bankers, loans, interest
	Banking industry	firm, investment, banking, business, head, management, securities, group, financial, chief, trading, global, capital, chairman
	Insurance industry	insurance, company, life, companies, policies, coverage, insurers, insurer, policy, percent, first, claims, premiums, business
	Real estate industry	estate, real, building, property, office, square, construction, development, center, project, space, buildings, new, park
	Consultancy	firm, clients, partner, partners, firms, law, client, lawyers, consulting, lawyer, consultants, business, former, help

(continued)

**Appendix.** (continued)

Semantic Groups	Semantic Patterns	Top Words
Pharmacy and health care	Pharmacy	drug, products, drugs, companies, company, pharmaceutical, product, health, generic, last, supplements, medical, testing, medicine
	Health care	health, care, medical, hospital, doctors, costs, hospitals, patients, medicare, insurance, coverage, system, cost, benefits
	Food retail	food, foods, company, restaurant, coffee, products, restaurants, brands, glass, chain, beverage, soft, grand, meat
	Merchandise and fashion	stores, store, retail, sales, retailers, department, chain, merchandise, retailer, apparel, goods, merchandising, new, fashion
	Textile industry	industry, cent, national, association, new, manufacturers, production, convention, textile, annual, trade, aluminum, consumer, use
Logistics	Retail	products, consumer, product, consumers, companies, percent, company, paper, new, customers, marketing, year, sales, customer
	Railroad	city, new, railroad, service, central, first, transportation, mayor, local, cities, area, line, hall, railroads
	Automotive	car, auto, cars, motors, general, new, motor, American, sales, year, plant, production, dealers, market
	Aviation	airlines, airline, air, travel, American, flight, continental, aviation, airport, eastern, international, pilots, carrier, passengers
		(continued)

# Appendix. (continued)

Semantic Groups	Semantic Patterns	Top Words
Heavy industry	Military industry	military, defense, space, government, contract, program, contracts, aircraft, contractors, force, air, contractor, navy, general
Agriculture	Steel industry	year, steel, quarter, percent, sales, last, first, earnings, loss, net, cents, income, share, increase
Public administration	Agriculture	farm, year, farmers, years, state, land, area, water, wood, river, old, lake, trees, grain
	Legislation	industry, congress, commission, new, government, federal, agency, law, legislation, rules, committee, administration, proposal, officials
	Administration and economy	public, system, government, private, role, interest, power, economic, issues, policy, society, term, issue, view
	Education	school, university, business, students, college, program, education, schools, training, job, high, professor, programs, student
International trade	International trade	European, foreign, international, countries, American, world, German, government, trade, country, British, French, global, Chinese

(continued)

**Appendix.** (continued)

Semantic Groups	Semantic Patterns	Top Words
Functional processes	Sales, costs, and prices	price, market, industry, prices, costs, sales, cost, competition, last, rates, pricing, high, demand, makers
	Sales and marketing	company, new, sales, division, group, marketing, business, operations, president, products, international, unit, changes, divisions
	Research and development	company, research, companies, equipment, products, manufacturing, engineering, technology, industrial, year, production, new, development, American
	Accounting and auditing	accounting, financial, audit, company, accountants, auditors, statements, fraud, companies, public, auditor, reports, new, securities
	Merger and acquisition	company, deal, merger, corporation, companies, agreement, group, sale, business, acquisition, stake, yesterday, percent, venture




## Acknowledgments

We are grateful to John W. Mohr and Peter Walgenbach for their very supportive comments on a previous draft of this manuscript. Moreover, we appreciate the help of Lisa-Maria Gerhardt, David Hoffmann, Marcus Rappe, and Karoline Wolkersdorfer during the data preparation process, and we thank Johannes A. Schubert for his support during the software development process. Finally, we thank our editor Duane F. Alwin and three anonymous reviewers for their great support and recommendations during the review process.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Deutsche Forschungsgemeinschaft (grant no. WA 2139/16-1).

## ORCID iD

Philipp Poschmann  <https://orcid.org/0000-0002-7056-6013>

## Notes

1. We implemented the proposed methodical approach in a self-developed software application. This software is available for free (<https://jenpara.uni-jena.de>). For natural language processing, we draw on state-of-the-art software, namely, Stanford's CoreNLP (Manning et al. 2014) and the latent Dirichlet allocation algorithm (Blei et al. 2003) implemented in the "lda" Python package. Our software implements and provides the most basic functions needed for extracting semantic triplets and detecting semantic patterns in texts.
2. Note that the data sources, which were compiled to a text corpus, may influence the results. Therefore, our proposed formal in-depth analysis does not free social scientists from carefully considering the fit of text corpus and research questions. Social scientists should consider which layers of meaning they wish to investigate and consequently, whether different data sources may influence their results. For example, with communication structures and their semantic surroundings, our formal in-depth analysis builds on layers of meaning that do not depend heavily on pragmatics, namely, questions of how to do things with words (cf. Austin 1975). With regard to communication structures, we statistically checked for the time span 1995 to 2005, which we randomly selected, whether the *New York Times* and the *Washington Post* used similar words in certain grammatical positions. We found that the *New York Times* and the *Washington Post* correlated, on average, at a level of .81. This illustrates that the different data sources used similar words when referring to responsibility, and thus having two sources does not extensively influence our results. However, this finding does not exclude the possibility that the similarity between data sources may be low on more subtle layers of meaning—for instance, on the level of pragmatics, which may be due to different ideological tendencies. In this

sense, by facilitating first interpretations based on a formal in-depth analysis, the strength of our approach is to enable and guide subsequent close reading. Such close reading may in turn capture the multiplexity of linguistic nuances across different data sources.

3. Our methodological framework also works with other software packages. For instance, grammatical parsing is included in OpenNLP (<https://opennlp.apache.org/>), NLTK (<https://www.nltk.org/>), and MaltParser (<http://www.maltparser.org/>). However, grammatical parsing still remains widely inaccessible to social scientists who are not familiar with computer science and software development because the extraction of semantic triplets requires programming skills. Therefore, providing a fully developed and reliable software application that enables grammatical parsing and the extraction of semantic triplets is a rewarding endeavor.
4. Topic modeling is also included in Mallet (<http://mallet.cs.umass.edu/index.php>), topicmodels (<https://cran.r-project.org/web/packages/topicmodels/index.html>), and gensim (<https://radimrehurek.com/gensim/>). Notably, our software's output, which depicts semantic triplets, is suitable for researchers to add the topic modeling results from each software package they prefer. We include our own topic-modeling script to make topic modeling broadly accessible for social scientists. In addition, we provide the software script needed for a quantitative validity test of topic modeling.

## References

- Abell, Peter. 1987. *The Syntax of Social Life: The Theory and Method of Comparative Narratives*. Oxford, UK: Clarendon Press.
- Abend, Gabriel. 2014. *The Moral Background: An Inquiry into the History of Business Ethics*. Princeton, NJ: Princeton University Press.
- Abrahamson, Eric, and Micki Eisenman. 2008. "Employee-Management Techniques: Transient Fads or Trending Fashions?" *Administrative Science Quarterly* 53(4): 719–44.
- Abramson, Corey M., Jacqueline Joslyn, Katharine A. Rendle, Sarah B. Garrett, and Daniel Dohan. 2018. "The Promises of Computational Ethnography: Improving Transparency, Replicability, and Validity for Realist Approaches to Ethnographic Analysis." *Ethnography* 19(2):254–84.
- Allison, Sarah, Ryan Heuser, Matthew L. Jockers, Franco Moretti, and Michael Witmore. 2011. *Quantitative Formalism: An Experiment*. Stanford, CA: Stanford University Press.
- Amable, Bruno. 2011. "Morals and Politics in the Ideology of Neo-liberalism." *Socio-Economic Review* 9(1):3–30.
- Austin, John L. 1975. *How to Do Things with Words*. 2nd ed. Cambridge, MA: Harvard University Press.
- Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43(3):465–524.
- Baumer, Eric P. S., David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. "Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence?" *Journal of the Association for Information Science and Technology* 68(6):1397–410.

- Berger, Peter L., and Thomas Luckmann. 1967. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. London: Penguin Press.
- Biber, Douglas, Susann Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge, UK: Cambridge University Press.
- Biernacki, Richard. 2009. "After Quantitative Sociology: Interpretive Science as a Calling." Pp. 119–208 in *Meaning and Method*, edited by I. Reed and J. C. Alexander. Boulder, CO: Paradigm Publishers.
- Biernacki, Richard. 2012. *Reinventing Evidence in Social Inquiry*. New York: Palgrave Macmillan.
- Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4): 77–84.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.
- Bonilla, Tabitha, and Justin Grimmer. 2013. "Elevated Threat Levels and Decreased Expectations: How Democracy Handles Terrorist Threats." *Poetics* 41(6):650–69.
- Boyer, Robert, and Joseph R. Hollingsworth. 1998. "From National Embeddedness to Spatial and Institutional Nestedness." Pp. 433–84 in *Contemporary Capitalism: The Embeddedness of Institutions*, edited by R. Boyer and J. R. Hollingsworth. Cambridge, UK: Cambridge University Press.
- Brammer, Stephen, Gregory Jackson, and Dirk Matten. 2012. "Corporate Social Responsibility and Institutional Theory: New Perspectives on Private Governance." *Socio-Economic Review* 10(1):3–28.
- Breiger, Ronald L., Robin Wagner-Pacifici, and John W. Mohr. 2018. "Capturing Distinctions While Mining Text Data: Toward Low-Tech Formalization for Text Analysis." *Poetics* 68(1):104–19.
- Brill, Eric. 2000. "Part of Speech Tagging." Pp. 403–14 in *Handbook of Natural Language Processing*, edited by R. Dale, H. Moisl, and H. Somers. New York: Marcel Dekker.
- Campbell, John L. 2007. "Why Would Corporations Behave in Socially Responsible Ways? An Institutional Theory of Corporate Social Responsibility." *Academy of Management Review* 32(3):946–67.
- Carley, Kathleen M. 1993. "Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis." *Sociological Methodology* 23(1):75–126.
- Carley, Kathleen M. 1996. "Artificial Intelligence within Sociology." *Sociological Methods & Research* 25(1):3–30.
- Carley, Kathleen M. 1997. "Extracting Team Mental Models through Textual Analysis." *Journal of Organizational Behavior Management* 18(S1):533–58.
- Carroll, John A. 2004. "Parsing." Pp. 233–48 in *The Oxford Handbook of Computational Linguistics*, edited by R. Mitkov. Oxford, UK: Oxford University Press.
- Carstensen, Kai-Uwe, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde, and Hagen Langer. 2010. *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Heidelberg: Spektrum Akademischer Verlag Heidelberg.

- Chakrabarti, Parijat, and Margaret Frye. 2017. "A Mixed-Methods Framework for Analyzing Text Data: Integrating Computational Techniques with Qualitative Methods in Demogra." *Demographic Research* 37(42):1351–82.
- Chen, Danqi, and Christopher D. Manning. 2014. "A Fast and Accurate Dependency Parser Using Neural Networks." Pp. 740–50 in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics.
- Cornelissen, Joep P., Rodolphe Durand, Peer C. Fiss, John C. Lammers, and Eero Vaara. 2015. "Putting Communication Front and Center in Institutional Theory and Analysis." *Academy of Management Review* 40(1):10–27.
- Cornelissen, Joep P., and Mirjam D. Werner. 2014. "Putting Framing in Perspective: A Review of Framing and Frame Analysis across the Management and Organizational Literature." *Academy of Management Annals* 8(1):181–235.
- Davis, Gerald F. 2009. *Managed by the Markets: How Finance Re-shaped America*. Oxford, UK: Oxford University Press.
- de George, Richard T. 1999. *Business Ethics*. 5th ed. Upper Saddle River, NJ: Prentice Hall Press.
- de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. "Generating Typed Dependency Parses from Phrase Structure Parses." Pp. 449–54 in *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy: European Language Resources Association.
- DiMaggio, Paul J., Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41(6):570–606.
- Dobbin, Frank R., and Jiwook Jung. 2010. "The Misapplication of Mr. Michael Jensen: How Agency Theory Brought Down the Economy and Why It Might Again." Pp. 29–64 in *Markets on Trial: The Economic Sociology of the U.S. Financial Crisis*, edited by M. Lounsbury and P. M. Hirsch. Bingley, UK: Emerald.
- Edelmann, Achim, and John W. Mohr. 2018. "Formal Studies of Culture: Issues, Challenges, and Current Trends." *Poetics* 68(1):1–9.
- Eder, Maciej. 2014. "Stylometry, Network Analysis and Latin Literature." Pp. 457–58 in *Digital Humanities 2014: Book of Abstracts*. Lausanne: EPFL-UNIL.
- Evans, James A., and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42(1):21–50.
- Fillmore, Charles J. 1985. "Frames and the Semantics of Understanding." *Quaderni Di Semantica* 6(2):222–54.
- Finkel, Jr., Trond Grenager, and Christopher Manning. 2005. "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling." Pp. 363–70 in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics-ACL 2005*. Stroudsburg, PA: Association for Computational Linguistics.
- Fligstein, Neil, Jonah S. Brundage, and Michael Schultz. 2017. "Seeing Like the Fed: Culture, Cognition, and Framing in the Failure to Anticipate the Financial Crisis of 2008." *American Sociological Review* 82(5):879–909.
- Florian, Radu, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. "Named Entity Recognition through Classifier Combination." *Proceedings of the 7th Conference on Natural Language Learning (HLT-NAACL)* 4:168–71.

- Franzosi, Roberto. 1989. "From Words to Numbers: A Generalized and Linguistics-Based Coding Procedure for Collecting Textual Data." *Sociological Methodology* 19(1):263–98.
- Franzosi, Roberto. 1990. "Computer-Assisted Coding of Textual Data: An Application to Semantic Grammars." *Sociological Methods & Research* 19(2):225–57.
- Franzosi, Roberto. 2010. *Quantitative Narrative Analysis*. Los Angeles: Sage Publications.
- Franzosi, Roberto, Gianluca de Fazio, and Stefania Vicari. 2012. "Ways of Measuring Agency: An Application of Quantitative Narrative Analysis to Lynchings in Georgia (1875–1930)." *Sociological Methodology* 42(1):1–42.
- Ge, Niyu, John Hale, and Eugene Charniak. 1998. "A Statistical Approach to Anaphora Resolution." Pp. 161–70 in *Proceedings of the 6th Workshop on Very Large Corpora*. Montreal, Canada: The Association for Computational Linguistics.
- George, Gerard, Ernst C. Osinga, Dovev Lavie, and Brent A. Scott. 2016. "Big Data and Data Science Methods for Management Research." *Academy of Management Journal* 59(5):1493–507.
- Gildea, Daniel, and Daniel Jurafsky. 2002. "Automatic Labeling of Semantic Roles." *Computational Linguistics* 28(3):245–88.
- Goldenstein, Jan, Philipp Poschmann, Sebastian G. M. Händschke, and Peter Walgenbach. 2019. "Global and Local Orientation in Organizational Actorhood: A Comparative Study of Large Corporations from Germany, the United Kingdom, and the United States." *European Journal of Cultural and Political Sociology* 6(2): 201–36.
- Griffiths, Thomas L., and Mark Steyvers. 2004. "Finding Scientific Topics." Pp. 5228–35 in *Arthur M. Sackler Colloquium of the National Academy of Sciences*. Washington, DC: National Academy of Sciences.
- Hiß, Stefanie. 2009. "From Implicit to Explicit Corporate Social Responsibility: Institutional Change as a Fight for Myths." *Business Ethics Quarterly* 19(3):433–52.
- Jockers, Matthew L., and David Mimno. 2013. "Significant Themes in 19th-Century Literature." *Poetics* 41(6):750–69.
- Kabanoff, Boris, Robert Waldersee, and Marcus Cohen. 1995. "Espoused Values and Organizational Change Themes." *Academy of Management Journal* 38(4):1075–104.
- Kahl, Steven J., and Stine Grodal. 2016. "Discursive Strategies and Radical Technological Change: Multilevel Discourse Analysis of the Early Computer (1947–1958)." *Strategic Management Journal* 37(1):149–66.
- Kang, Nahee, and Jeremy Moon. 2012. "Institutional Complementarity between Corporate Governance and Corporate Social Responsibility: A Comparative Institutional Analysis of Three Capitalisms." *Socio-Economic Review* 10(1):85–108.
- Kaplan, Rami. 2015. "Who Has Been Regulating Whom, Business or Society? The Mid-20th-century Institutionalization of 'Corporate Responsibility' in the USA." *Socio-Economic Review* 13(1):125–55.
- Kennedy, Mark T. 2008. "Getting Counted: Markets, Media, and Reality." *American Sociological Review* 73(2):270–95.
- Khaire, Mukti, and R. Daniel Wadhvani. 2010. "Changing Landscapes: The Construction of Meaning and Value in a New Market Category—Modern Indian Art." *Academy of Management Journal* 53(6):1281–304.

- Kirchner, Corinne, and John W. Mohr. 2010. "Meanings and Relations: An Introduction to the Study of Language, Discourse and Networks." *Poetics* 38(6):555–66.
- Kobourov, Stephen G. 2012. "Spring Embedders and Force Directed Graph Drawing Algorithms." *ArXiv* (1201.3011):1–23.
- Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Thousand Oaks, CA: Sage Publications.
- Langacker, Ronald W. 1986. "An Introduction to Cognitive Grammar." *Cognitive Science* 10(1):1–40.
- Langley, Pat. 1996. *Elements of Machine Learning*. San Francisco: Morgan Kaufmann.
- Lasswell, Harold D., Daniel Lerner, and Ithiel de Sola Pool. 1952. *The Comparative Study of Symbols. An Introduction*. Palo Alto, CA: Stanford University Press.
- Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Daniel Jurafsky. 2013. "Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules." *Computational Linguistics* 39(4):885–916.
- Lee, Monica, and John Levi Martin. 2015. "Coding, Counting and Cultural Cartography." *American Journal of Cultural Sociology* 3(1):1–33.
- Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, and Christian Bizer. 2015. "DBpedia—A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia." *Semantic Web* 6(2):167–95.
- Maignan, Isabelle, and David A. Ralston. 2002. "Corporate Social Responsibility in Europe and the U.S.: Insights from Businesses' Self-Presentations." *Journal of International Business Studies* 33(3):497–514.
- Malsberger, John W., and James N. Marshall, eds. 2009. *The American Economic History Reader*. New York: Routledge.
- Manning, Christopher D., John Bauer, Jenny Finkel, Steven J. Bethard, Mihai Surdeanu, and David McClosky. 2014. "The Stanford CoreNLP Natural Language Processing Toolkit." Pp. 55–60 in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Stroudsburg, PA: Association for Computational Linguistics.
- Manning, Christopher D., and Hinrich Schütze. 2000. *Foundations of Natural Language Processing*. 4th ed. Cambridge, MA: MIT Press.
- Marens, Richard. 2012. "Generous in Victory? American Managerial Autonomy, Labour Relations and the Invention of Corporate Social Responsibility." *Socio-Economic Review* 10(1):59–84.
- Matten, Dirk, and Jeremy Moon. 2008. "'Implicit' and 'Explicit' CSR: A Conceptual Framework for a Comparative Understanding of Corporate Social Responsibility." *Academy of Management Review* 33(2):404–24.
- Matthews, Peter H. 1981. *Syntax*. Cambridge, UK: Cambridge University Press.
- McCormick, Tyler H., Hedwig Lee, Nina Cesare, Ali Shojaie, and Emma S. Spiro. 2017. "Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing." *Sociological Methods & Research* 46(3):390–421.
- McWilliams, Abigail, and Donald Siegel. 2001. "Corporate Social Responsibility: A Theory of the Firm Perspective." *Academy of Management Review* 26(1):117–27.
- Mel'čuk, Igor A. 1988. *Dependency Syntax: Theory and Practice*. Albany: SUNY Press.

- Meyer, John W., Shawn Pope, and Andrew Isaacson. 2015. "Legitimizing the Transnational Corporation in a Stateless World Society." Pp. 27–72 in *Corporate Social Responsibility in a Globalizing World*, edited by K. Tsutsui and A. Lim. Cambridge, UK: Cambridge University Press.
- Meyer, Renate E., and Markus A. Höllerer. 2010. "Meaning Structures in a Contested Issue Field: A Topographic Map of Shareholder Value in Austria." *Academy of Management Journal* 53(6):1241–62.
- Mitchell, Tom. 1997. *Machine Learning*. New York: McGraw-Hill Book.
- Mitkov, Ruslan, ed. 2004. "Anaphora Resolution." Pp. 266–83 in *The Oxford Handbook of Computational Linguistics*. Oxford, UK: Oxford University Press.
- Mohr, John W. 1998. "Measuring Meaning Structures." *Annual Review of Sociology* 24(1):345–70.
- Mohr, John W., and Petko Bogdanov. 2013. "Introduction Topic Models: What They Are and Why They Matter." *Poetics* 41(6):545–69.
- Mohr, John W., Robin Wagner-Pacifici, and Ronald L. Breiger. 2015. "Toward a Computational Hermeneutics." *Big Data & Society* 2(2):1–8.
- Mohr, John W., Robin Wagner-Pacifici, Ronald L. Breiger, and Petko Bogdanov. 2013. "Graphing the Grammar of Motives in National Security Strategies: Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics." *Poetics* 41(6):670–700.
- Moretti, Franco. 2013. *Distant Reading*. London: Verso.
- Muller, Michael, Shion Guha, Eric P. S. Baumer, David Mimno, and N. Sadat Shami. 2016. "Machine Learning and Grounded Theory Method." Pp. 3–8 in *Proceedings of the 19th International Conference on Supporting Group Work—GROUP '16*. New York, NY: Association for Computing Machinery.
- Nardulli, Peter F., Scott L. Althaus, and Matthew Hayes. 2015. "A Progressive Supervised-Learning Approach to Generating Rich Civil Strife Data." *Sociological Methodology* 45(1):148–83.
- Nederhof, Mark-Jan, and Giorgia Satta. 2013. "Theory of Parsing." Pp. 105–30 in *The Handbook of Computational Linguistics and Natural Language Processing*, edited by A. Clark, C. Fox, and S. Lappin. New York: John Wiley & Sons.
- Nelson, Laura K. 2017. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research*. doi:10.1177/0049124117729703.
- Palmer, Ian, Boris Kabanoff, and Richard Dunford. 1997. "Managerial Accounts of Downsizing." *Journal of Organizational Behavior* 18(1):623–39.
- Pedersen, Ove K. 2010. "Institutional Competitiveness: How Nations Came to Compete." Pp. 625–58 in *The Oxford Handbook of Comparative Institutional Analysis*, edited by G. Morgan, J. L. Campbell, C. Crouch, O. K. Pedersen, and R. Witley. Oxford, UK: Oxford University Press.
- Phillips, Nelson, Thomas B. Lawrence, and Cynthia Hardy. 2004. "Discourse and Institutions." *Academy of Management Review* 29(4):636–52.
- Pool, Ithiel de Sola. 1959. *Trends in Content Analysis*. Urbana: University of Illinois Press.
- Popping, Roel. 2012. "Qualitative Decisions in Quantitative Text Analysis Research." *Sociological Methodology* 42(1):88–90.

- Popping, Roel, and Carl W. Roberts. 2015. "Semantic Text Analysis and the Measurement of Ideological Developments within Fledgling Democracies." *Social Science Information* 54(1):23–37.
- Punyakanok, Vasin, Dan Roth, and Wen-tau Yih. 2008. "The Importance of Syntactic Parsing and Inference in Semantic Role Labeling." *Computational Linguistics* 34(2): 257–87.
- Roberts, Carl W. 1989. "Other Than Counting Words: A Linguistic Approach to Content Analysis." *Social Forces* 68(1):147–77.
- Roberts, Carl W. 2000. "A Conceptual Framework for Quantitative Text Analysis: On Joining Probabilities and Substantive Inferences about Texts." *Quality & Quantity* 34(3):259–74.
- Roberts, Carl W., Cornelia Zuell, Juliane Landmann, and Yong Wang. 2010. "Modality Analysis: A Semantic Grammar for Imputations of Intentionality in Texts." *Quality and Quantity* 44(2):239–57.
- Rule, Alix, Jean-Philippe Cointet, and Peter S. Bearman. 2015. "Lexical Shifts, Substantive Changes, and Continuity in State of the Union Discourse, 1790–2014." *Proceedings of the National Academy of Sciences* 112(35):10837–44.
- Santorini, Beatrice. 1990. *Part of Speech Tagging Guidelines for the Penn Treebank Project*. Philadelphia: University of Pennsylvania.
- Shamir, Ronen. 2011. "Socially Responsible Private Regulation: World-Culture or World-Capitalism?" *Law and Society Review* 45(2):313–36.
- Short, Jeremy C., J. Christian Broberg, and Keith H. Brigham. 2010. "Construct Validation Using Computer-Aided Text Analysis (CATA)." *Organizational Research Methods* 13(2):320–47.
- Simmons, Beth A., Frank Dobbin, and Geoffrey Garrett. 2006. "Introduction: The International Diffusion of Liberalism." *International Organization* 60(4):781–810.
- Streeck, Wolfgang. 2014. "How Will Capitalism End?" *New Left Review* 87(1):35–64.
- Sudhahar, Saatviga, Gianluca de Fazio, Roberto Franzosi, and Nello Cristianini. 2013. "Network Analysis of Narrative Content in Large Corpora." *Natural Language Engineering* 21(1):81–112.
- Sutton, Charles, and Andrew McCallum. 2006. "An Introduction to Conditional Random Fields for Relational Learning." Pp. 1–35 in *Introduction to Statistical Relational Learning*, edited by L. Getoor and B. Taskar. Cambridge, UK: MIT Press.
- Toutanova, Kristina, Dan Klein, and Christopher D. Manning. 2003. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." Pp. 252–59 in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*. Stroudsburg, PA: Association for Computational Linguistics.
- Trost, Harald. 2004. "Morphology." Pp. 25–47 in *The Oxford Handbook of Computational Linguistics*, edited by R. Mitkov. Oxford, UK: Oxford University Press.
- Tsur, Oren, Dan Calacci, and David Lazer. 2015. "A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns." Pp. 1629–38 in *Proceedings of the 53rd Annual Meeting of the Association for Computational*



- Linguistics and the 7th International Joint Conference on Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics.
- van Attevelde, Wouter, Jan Kleinnijenhuis, and Nel Ruigrok. 2008. "Parsing, Semantic Networks, and Political Authority: Using Syntactic Analysis to Extract Semantic Relations from Dutch Newspaper Articles." *Political Analysis* 16(4):428–46.
- Vicari, Stefania. 2010. "Measuring Collective Action Frames: A Linguistic Approach to Frame Analysis." *Poetics* 38(5):504–25.
- Vidal, Matt. 2011. "Reworking Postfordism: Labor Process Versus Employment Relations." *Sociology Compass* 5(4):273–86.
- Vidal, Matt. 2013. "Postfordism as a Dysfunctional Accumulation Regime: A Comparative Analysis of the USA, the UK and Germany." *Work, Employment and Society* 27(3):451–71.
- Vossen, Piek. 2004. "Ontologies." Pp. 464–82 in *The Oxford Handbook of Computational Linguistics*, edited by R. Mitkov. Oxford, UK: Oxford University Press.
- Voutilainen, Atro. 2004. "Part of Speech Tagging." Pp. 219–32 in *The Oxford Handbook of Computational Linguistics*, edited by R. Mitkov. Oxford, UK: Oxford University Press.
- Wagner-Pacifici, Robin, John W. Mohr, and Ronald L. Breiger. 2015. "Ontologies, Methodologies, and New Uses of Big Data in the Social and Cultural Sciences." *Big Data & Society* 2(2):1–11.
- Wasserman, Stanley, and Katherine Faust. 1999. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.

### Author Biographies

**Jan Goldenstein** is a research assistant at Friedrich Schiller University Jena in Germany. His research interests include institutional theory, linguistics, and natural language processing.

**Philipp Poschmann** is a research assistant at Friedrich Schiller University Jena in Germany. His research interests include social field theory, computer science, and computational linguistics.