Journal for the Theory of Social Behaviour 46:1
DOI: 10.1111/jtsb.12086

Theoretical Foundations for Digital Text Analysis

GABE IGNATOW

ABSTRACT

Much of social life now takes place online, and records of online social interactions are available for social science research in the form of massive digital text archives. But cultural social science has contributed little to the development of machine-assisted text analysis methods. As a result few text analysis methods have been developed that link digital text data to theories about culture and discourse. This paper attempts to lay the groundwork for development of such methods by proposing metatheoretical and theoretical foundations suitable for machine-assisted semantic text analysis. Metatheoretically I draw on the work of Elder-Vass (2012), Kaidesoja (2013) and others to argue that digital text analysis methods ought to be (and in practice implicitly are) based on a realist constructionist ontology that treats discourses as ontologically real emergent social entities that have causal relationships with non-discursive social and cognitive processes. Theoretically I follow Feldman (2006) and many others in arguing that language is fundamentally shaped by processes of embodied cognition. Researchers developing digital text analysis techniques must theoretically account for such processes if they wish to produce algorithms that can interpret texts in ways that supplement, and not only amplify, human interpretation. I critically survey contemporary text analysis methods that implicitly share these metatheoretical and theoretical positions and discuss some ways these can be further developed with newly available software.

Keywords: text analysis, philosophy of social science, theory of language, digital social science, big data.

I. INTRODUCTION

In the social sciences text mining and analysis methods have developed in a haphazard manner (Ruiz Ruiz, 2009; Bauer, Bicquelet & Suerdem, 2014). The small number of social scientists who work with such methods often list them as secondary or tertiary research interests. There are few social science institutes,

fellowships, or summer workshops dedicated to text mining or to quantitative and mixed qualitative-quantitative text analysis methods. While social science methods textbooks often cover qualitative content analysis and discourse analysis methods, they rarely cover quantitative or mixed methods for analyzing large text collections.

Today with ICTs (information communication technologies), social media, digital archives and the "big data" revolution, there is a pressing need for the social sciences to get their house in order with regard to the development and application of machine-assisted text analysis methods (Bail, 2014). Such methods are widely used in computational linguistics and marketing applications, and computer scientists and linguists are developing new text analysis techniques and software packages at a rapid pace. But even as many social researchers are gravitating toward digital text mining and analysis methods, a number of critical metatheoretical and theoretical issues remain unresolved or unaddressed.

In terms of metatheory (and philosophy of social science more generally), social researchers have not systematically articulated what large digital text collections can potentially reveal about the groups and communities that create them, how machine-assisted analysis of large digital text collections can improve on human interpretation of smaller text samples, or the limits of what can be learned from digital methods. This is particularly problematic for social research concerned with culture and discourse. Given that texts are open to multiple interpretations that depend on complex background knowledge and subtle contextual cues, how can software possibly tell us anything about intersubjective meaning creation occurring within social groups that we cannot learn from interpretive methods such as conversation analysis or discourse analysis?

In terms of theory, social researchers who develop and use digital text analysis methods have rarely specified the theory of language upon which their methods rest. And where they have specified such theories they have relied on theories of language that have been shown to be inadequate for the task of modeling intersubjective meaning construction.

In light of the increasing availability of large text data sets and interest in machine-assisted text analysis methods, the goal of the present paper is to provide clarity or at a minimum generate discussion as to what these techniques can potentially contribute to social science research concerned with culture and discourse. I attempt to lay the groundwork for development of social research methods that link digital text data to theoretical debates over cultural and discursive social processes by proposing metatheoretical and theoretical foundations suitable for machine-assisted semantic text analysis methods.

I have two metatheoretical starting points: Elder-Vass's (2010, 2012) "realist constructionist" ontology, and Kaidesoja's (2013) closely related argument for a naturalized realist ontology. With others (e.g. Fairclough, Jessop, & Sayer, 2007; Sealey, 2009), Elder-Vass has argued that discourses ought to be understood as ontologically real emergent social entities, and that in their spontaneous practices

many researchers "implicitly adopt realist assumptions" (Elder-Vass, 2013, p. 250) even when they may take an anti-realist, constructionist stance in their metascientific refelctions. Kaidesoja's (2013) argument is that a naturalized realist ontology implies that discourses are amenable to rigorous formal analysis to roughly the same degree as any other social phenomenon. While a naturalized realist constructionist ontology accords with the spontaneous research practices and common sense understandings of many social researchers, explicating an ontology that is implicit in most machine-assisted text analysis research is a useful exercise. It can help to distance such text analysis research from research founded on extreme constructionist ontologies that are philosophically untenable and strategically counterproductive. And it provides a philosophical foundation for the claim that interpretations of human language produced by quantitative text analysis methods can in some cases be superior to interpretations produced by exegetical methods.

Theoretically, I follow Feldman (2006) and many others (Bergen & Chang, 2005; Goldberg, 1995; Lakoff, 2012) in arguing that language is fundamentally shaped by processes of embodied cognition. Language is experienced as meaningful because it is inseparable from human bodily capacities and operations. Because language simulates rather than encodes embodied experience, formal models of language that do not incorporate bodily operations, including emotions, sensations, and perceptions, are highly artificial. The implication here is that researchers who use formal semantic text analysis methods should theoretically account for processes of embodied cognition if they wish to produce algorithms capable of revealing intersubjective meaning construction from large text collections.

After advocating specific metatheoretical and theoretical foundations for development of machine-assisted semantic text analysis methods, I critically survey text analysis methods that implicitly share some or all of these metatheoretical and theoretical presuppositions. I argue that greater awareness of the metatheoretical and theoretical foundations of text analysis methods can allow researchers to more precisely position machine-assisted text analysis research relative to both theoretically oriented exegetical text analysis techniques and inductive text mining techniques. I also suggest some ways semantic text analysis methods can be further developed by incorporating rapidly developing methods for sentiment analysis. Semantic text analysis methods constructed on the metatheoretical and theoretical foundations proposed in this paper can potentially provide social researchers with new tools for exploring cultural differences, tracking and predicting changes in public mood, and predicting behavior on- and off-line.

II. THE DIGITAL TEXT REVOLUTION AND AUTOMATED TEXT ANALYSIS

While sociologists have recently begun to use archived digital text data for social network analysis (Lewis, Gonzalez, & Kaufman, 2012), analysis of trends in media

coverage of contentious issues (Ignatow & Williams, 2011), and text analysis of sites and documents produced by social movements (Caren, Jowers, & Gaby, 2012), they have only begun to explore the implications of digitally archived textual data for developing and refining theory (Bail, 2014; Mohr & Rawlings, 2012).

While qualitative content analysis is commonly used in sociology for analyzing ethnographic interview transcripts, open-ended survey item responses, newspapers, and social media texts, quantitative and mixed-method text analysis techniques are less common. Roberts (1997) has categorized such quantitative techniques as either *thematic*, *network*, or *semantic* techniques. Thematic text analysis techniques focus on manifest meanings in texts, and include methods commonly used in business as well as social science, such as topic modeling (Mohr & Bogdanov, 2013). Network text analysis methods model statistical associations between words to infer the existence of mental models shared by members of a community (Carley, 1997). Semantic text analysis methods (Mohr & Rawlings [2010] refer to these as hermeneutic or hermeneutic structuralist approaches) include a variety of methods designed to recognize latent meanings in texts (e.g. Franzosi, 2010; Gottschalk & Gleser, 1969; Ignatow, 2004, 2009; Roberts, 1997; Schrodt & Savaiano, 1997). The purpose of the present paper is to advocate specific metatheoretical and linguistic-theoretical foundations for further development of semantic text analysis methods for use on large digital text collections.

III. METATHEORY: CONSTRUCTIONIST AND NATURALIST REALISMS

Social researchers who work with machine-assisted text mining and analysis methods have not often engaged with contemporary philosophy of social science, and recent debates over metatheory among philosophers of social science appear to many empirical researchers to be of little direct relevance to their work. But there are at least two reasons why researchers using quantitative and mixed text analysis methods ought to give some consideration to the metatheoretical positions that are implied in their methodological approaches and to how their methods link to theory. First, digital text analysis techniques are developing rapidly but haphazardly (Ruiz Ruiz, 2009). Articulating what these techniques are potentially capable of, and why, may allow interested researchers to better understand how automated methods may improve upon more well established research paradigms and methods, including especially exegetical methods. Second, because semantic quantitative text analysis techniques are uniquely positioned at the intersection of the "two cultures" of the sciences and humanities, the ontological underpinnings of such techniques are uniquely unsettled. Semantic text analysis research may be done in a realist-positivist epistemic mode that values quantification, hypothesis testing, and statistical analysis; in an interpretive mode that values close reading and multiple interpretations of texts; or, most often, in an ad hoc combination of these two modes. That these different modes of conducting

research often produce incommensurable findings is widely recognized and debated in the social sciences (Reed, 2011; Biernacki, 2012). Metatheoretical reflection and critique can clarify where studies using quantitative and automated text mining and analysis tools will necessarily diverge from, and potentially have advantages and disadvantages relative to, studies that rely on purely interpretive methods, such as the neo-Durkheimian school of hermeneutic cultural sociology (Reed, 2008; Reed & Alexander, 2009) or the work of contemporary ethnographers (Pugh, 2009) and constructionist theorists (Zerubavel, 2009).

Social scientists who work with machine-assisted text mining and analysis methods are well aware that meanings of texts are always open to multiple interpretations. Both Elder-Vass (2012) and Mohr and Rawlings (2012) note that after social constructionism, postmodernism and the cultural turn there are few "naive realists" (Elder-Vass, 2012) left in even the natural and physical sciences. "[S]ome version of social constructionism" is a "core element defining the theoretical background" of most social scientific fields today" (Mohr & Rawlings, 2012). Yet while social researchers who use software and programming languages for text mining and analysis are not naive realists, in the spontaneous practice of research they generally operate in a realist "epistemic mode" (Reed, 2011), implicitly treating discourses as real social entities with qualities and causal powers that are discoverable through the development of appropriate research designs and application of rigorous methods and statistical techniques. Mohr and Rawlings refer to a realist "post-cultural turn sensibility" that views culture as a "force that shapes the social" (2012, p. 124). Kaufman (2004) has explicated this realist understanding of discourses and culture as entities that have causal relationships with exogenous factors (markets, organizational structures, social networks, cognitive mechanisms and other factors).

In philosophy of social science terms, the "post-cultural turn sensibility" implicit in quantitative and mixed methods text analytic research most closely resembles recently developed realist ontological positions including, *inter alia* (Fairclough, Jessop, & Sayer, 2007), Elder-Vass's realist constructionism (2012) and Kaidesoja's naturalist realism (2013). Elder-Vass's version of critical realism, his "realist constructionist" ontology, combines moderate constructionism and moderate realism. He argues that realist constructionism is superior to naive realism (the idea, now rarely held, that language and cognition produce universally valid reflections of an objective world) and extreme constructionism (the idea that there is nothing objectively or universally true, but rather that all truths are produced by members of human communities arbitrarily agreeing to refer to them as true).

Kaidesoja's recent contribution to philosophy of social science is his argument for a naturalized realist ontology. His naturalized realist social ontology assumes that social phenomena are a part of nature, and thus theories within a naturalist social ontology are expected to be compatible with the well-established (epistemically successful) assumptions and presuppositions of natural and physical

sciences. Against transendental and *a prioiri* epistemological and ontological reflection, he argues that just as for theories in the natural sciences, naturalist social ontology should be built "by means of a posteriori arguments that take the epistemically successful scientific practices and well-confirmed results of different sciences as their premises . . . The relationship between theories developed in emprical sciences and in naturalist ontology should be seen as continuous" (2013, p. 203). Kaidesoja finds in cognitive science and related empirical fields resources that can be integrated into existing theoretical and methodological frameworks in the social sciences on the basis of a naturalist realist ontology. Such an integration, he suggests, can pay dividends in the study of language in social life, which

> should not be separated from cognitive processing of individuals and distributed cognitive systems. The emergence, maintenance and changes of linguistic meanings should rather be related to the cognitive activities of communicatively interacting individuals in specific material and cultural environments. This sort of pragmatic approach to language has recently been advanced in the field of cognitive linguistics . . . [a discipline which] provides useful conceptual resources for developing a naturalized social ontology of language that enables us to focus on the dynamical role of language in social life (2013, p. 204).

Both constructionist and naturalist realist ontologies provide conceptual resources for the social sciences to transcend dead-end philosophical debates over naive realism versus extreme constructionism. They also provide philosophical foundations for building bridges from cultural social science to cognitive, behavioral, and computer science. However, while my position in this paper is that some version of realist philosophy of social science is implicit in the spontaneous practices of researchers working with text mining and analysis methods, and that further development of such methods can be accelerated by bringing these philosophical positions to light, there is at least one tension within realist philosophy of social science that is relevant to researchers using text mining and analysis methods. Theorists and empirical researchers who agree that culture is a real entity disagree about whether it emerges from social interactions as a collective representation that is irreducible to the subjective experiences of the individuals who produce it, or whether it exists within people's minds as mental representations but that the idea of some extra-individual cultural realm is fanciful (see Elder-Vass, 2012; Lizardo, 2007; Sperber, 1975; Turner, 2002). Bloch (1998), Sperber (1975), and Turner (2002) are among the more prominent advocates of a cognitivist and anti-emergentist approach to cultural analysis, Elder-Vass takes a middle ground with his concept of "linguistic norm circles," and Lizardo (2007) has suggested that neuroscience research on mirror neurons that shows that cultural transmission can occur below the level of conscious awareness supports a more sociological and emergentist position. While I happen to be most sympathetic to Elder-Vass's and Lizardo's positions, what I am advocating in this paper is the idea that text analysis methods can potentially inform theoretical debates concerning culture and discourse (just as they can contribute to many other social research areas) in

much the same way cognitive neuroscience has informed such debates over the last two decades (e.g. Sun, 2012; Turner, 2001), albeit primarily for audiences who are sympathetic to naturalism (and not committed to defending a strict divide between the human and natural sciences). Text mining and analysis methods have arguably even more potential relevance for cultural theory than does cognitive neuroscience because, as methods rather than bodies of research, they can be used in research projects designed to directly test middle-range theories derived from broad cultural theoretical positions (see Bail, 2014).

In the next section I discuss some specific contributions cognitive science, neuroscience, and computer science, all of which operate more or less implicitly within a naturalist and realist constructionist ontology, can make to the development of machine-assisted text analysis methods.

IV. THEORIES OF LANGUAGE: FOUR PSYCHOLINGUISTIC MODELS

Using software to interpret texts produced by social groups is necessarily a theory-driven enterprise because it requires a theory of language as a conceptual basis for developing scalable coding strategies. Thus far social scientists working with semantic text analysis methods have based their coding strategies on four main psycholinguistic theoretical models: Saussurian *binary structural models*, *sequence models*, *network models*, and *embodied cognitive models*. In what follows I review these four theoretical models and research based on each, and argue that embodied cognitive models of language are far better supported by contemporary research than are binary, sequential, or network models. Embodied cognitive models are the superior choice among available psycholinguistic models for use in semantic text analysis applications because they treat bodily and emotional capacities as central to and inseparable from intersubjective meaning construction. Theoretically my position resembles Shalin's recently developed sociological pragmatist hermeneutics (Shalin, 2007, p. 195). A methodological implication of this position is that semantic text analysis should move on from reductionist structuralist models, and efforts to identify "meaning structures" (Mohr, 1998) in texts should search for word-meaning associations by integrating sentiment analysis with methods for analyzing word frequencies and co-occurrences. Later in the paper I review several ways social scientists are beginning to accomplish this integration.

Binary Models

When sociologists have used large text corpora to analyze intersubjective meaning construction they have almost always relied on structuralist models of cognitive functioning. Structuralist models allow sociologists to reduce the complexity of language to elemental semantic units that because of their relative structural

simplicity are amenable to analysis with quantitative methods. Mohr and Rawlings (2012) refer to this mode of analysis as "hermeneutic structuralism," while Biernacki references a "structuralist orientation toward discourse" (2009). Although structural linguistics arguably began with Durkheim's sociological analysis of sacred and profane categories in religious life (Alexander, 1990a, pp. 4–5), since the 1980s the key figures for structuralism in cultural sociology have been Saussure and Levi-Strauss. Reading Durkheim's late work on religion through Saussure, Levi-Strauss, and the "linguistic turn, cultural sociologists in the neo-Durkheimian school seek to identify the binary "cultural codes" that are thought to structure social discourses. They have analyzed discourses ranging from media coverage of the 1970s Watergate scandal (Alexander, 1990b) to coverage of the Rodney King beating (Jacobs, 2000) and the Clinton/Lewinsky affair in the 1990s (Mast, 2006). Neo-Durkheimian researchers assert that cultural systems and the human mind are both organized in terms of oppositional binaries (sacred/profane, good/evil). But while the neo-Durkheimian school of cultural sociology contains many psychological and emotional elements, it also "brackets off naturalism in virtually all areas (McLennan, 2005, p. 10). While Neo-Durkheimians claim that psychology supports their understanding of generative binary psycholinguistic structures, they only occasionally reference psychoanalytic theory while scrupulously avoiding engagement with modern research psychology or cognitive neuroscience (cf. Ignatow, 2003). This stance has become increasingly problematic as the century-old Durkheimian/Saussurian binary psycholinguistic model has been superseded by newer models in nearly every academic discipline.

Sequential Models

A second set of structuralist psycholinguistic models are sequential models, also referred to as "narrative grammars" or "story grammars." Like binary models, sequential models attempt to capture a basic structural element of human cognition. In the case of sequential models that element is a fundamental social cognitive process whereby people interpret situations of all kinds in terms of basic social relations of actors, actions, and objects of action. Franzosi's term for these sequential structures is the "semantic triplet" or "S-A-O triplet." A pioneer in the use of narrative grammars for sociological text analysis (see also Cerulo, 1998), Franzosi has developed methods of sequential text analysis that involve teams of manual coders coding collections of historical texts, such as newspaper archives (1987), for S-A-O triplets. Franzosi and his collaborators have applied this method in studies of newspaper accounts of lynchings (Franzosi, De Fazio & Vicari, 2012) and of the rise of fascism (Franzosi, 2010), while Cerulo has used it in her studies of "victim" and "perpetrator" sequences in newspaper headlines (Cerulo, 1998), and Ignatow (2004) used sequence analysis in a multi-method study of transcripts of shipyard union leaders' meetings.

Sequential text analysis has traditionally required large teams of manual coders. However, recently Sudhahar, Franzosi, and Christianini (2011) have attempted to automate sequence analysis in order to take advantage of big data and inexpensive computer processing power. Yet even if it is eventually highly automated, the ability of sequence analysis to extend human interpretive capacities is limited simply because it does not incorporate bodily and emotional operations either theoretically or methodologically, and as we will see, bodily and emotional capacities have been shown to be fundamental to language and communication of all kinds.

Network Models

Rather than attempting to reduce the complexity of language by selecting from text samples only words related to binary or sequential structures, a number of social scientists who use text analysis methods model language in terms of networks of concepts. In the 1990s Carley (1994, 1997) developed "map analysis," which involves identifying concepts in texts and quantifying concepts' relationships to each other. Carley defines "social knowledge" as when concepts are shared by more than 50% of texts analyzed, "cultural diversity" as the number of concepts used in a text, and "cultural density" as the degree to which the social knowledge that forms the basis of a culture is interconnected. Carley and her colleagues have used this form of network-based text analysis in studies of software engineering teams (Carley, 1997) and of literature (Carley & Kaufer, 1993) and other topics (see Carley, 1994).

Network models are more complex than binary or sequential structuralist models, and like binary and sequential models they capture a basic structural characteristic of human cognition—in this case its associationist structure. But network models treat "concepts" as disembodied "mental models." Carley and Palmquist (1992, p. 602) state explicitly that their theoretical stance that "mental models are internal representations" underlies their methodology. As we will see, recent developments in cognitive neuroscience call into question the presupposition that interpretive and communicative practices can be understood in terms of purely "mental" structures without reference to bodily and emotional operations.

Embodied Cognitive Models

Following the cognitive revolution of the 1960s–70s and the establishment of cognitive science as an independent discipline, cognition came to be modeled as operating independently of bodily processes. For several decades it was widely assumed that cognition could be modeled without reference to the human body in much the same way that it is possible to understand the abstract logical operations

of computer code without any understanding of the hardware on which it runs. Since the 1980s this disembodied approach to cognition has been subject to sustained criticism on many fronts, partly because it never specified what cognitive processes actually are, how they shape action, or how they are associated with bodily and emotional capacities (Barsalou, 1999; Clark, 1997; Dreyfus, 2006; Lizardo, 2009). In response theories of embodied cognition have emerged (Ignatow, 2007). These theories posit that cognition shares systems with perception at all levels of analysis. Cognition is a bodily process because the brain is part of the body, and language is generated through processes of embodied cognition that involve sensory, perceptual, and affective systems to varying degrees.

Cognitive linguistics is the research area that applies theories of embodied cognition and knowledge to linguistic theory and research. Central to cognitive linguistics is the idea that cognition is carried out by neural circuitry, and that cognition is meaningful because of the way neural circuits are connected to the body. Language and abstract conceptualization are both understood to be embodied in this way (Lakoff, 2012).

Over the last decade cognitive linguistic theories have been formalized in what are known as construction-based grammars. These grammars are based on the idea that the most basic semantic unit is not a binary, sequential, or network structure, but rather a form-meaning pairing. Such pairings are known as "constructions" (Fillmore, Kay, & O'Connor, 1988; Goldberg, 1995; Langacker, 1999; Östman & Fried 2004). In one prominent construction-based formalization, Embodied Construction Grammar (ECG), researchers are developing computational simulations based on motor and perceptual schemas, construction grammars, and phonological, syntactic, and semantic constraints. ECG is intended to serve as a tool for linguistic analysis, and to support statistically based models of language acquisition and comprehension (see Lakoff [2012] for an overview).

Embodied cognition research, cognitive linguistics, and construction grammars provide new resources for sociologists interested in social construction phenomena. Of course research from these fields does not imply that it is wrong to use abstract concepts in analyzing a collection of texts; embodiment may matter more for some linguistic content than for others. Nonetheless, a number of researchers have developed automated or partially automated semantic text analysis methods that recognize that linguistic practices are fundamentally embodied, and that meaning construction can be modeled in terms of form-meaning pairings rather than in terms of binary, network, or other semantic structures. Some of these new methods are reviewed in the following section.

V. METHODS

Theories of embodied cognition support recent sociological critiques of reductionist structuralist approaches to large-scale text analysis (e.g. Biernacki,

2009, 2012). But rather than abandoning all efforts to develop automated semantic text analysis methods, the rapid development of the field of cognitive linguistics and of construction grammars suggest the possibility of evolving new text mining and analysis methods that model intersubjective meaning construction in a way that recognizes the central role of bodily and emotional operations. Sentiment analysis techniques are of central importance for such methods, and several families of text analysis methods have emerged recently that attempt to integrate sentiment analysis into semantic text analysis. These methods have been developed by both social scientists and computational linguists, often working collaboratively.

1. Human Coding

One method for integrating sentiment analysis into automated semantic text analysis is human coding of sentiment as expressed in texts. Bail has employed this method in a recent study of civil society organizations and media discourses on Islam post-September 11 (Bail, 2012). Human coding of sentiment in large text collections has the advantage of capturing cultural and contextual nuances in emotional expression that are often missed by highly automated methods. At the same time, human coding can be partially automated through application of supervised learning techniques. In supervised learning, human coders categorize a set of documents by hand, and then algorithms learn how to sort the remaining documents into categories using the training set of documents and words. Supervised methods thus require construction of a training set, application of the learning method which involves identifying relationships between features and categories in the training set and then using these to infer labels in the test set, and validating the model output and classifying remaining documents (see Grimmer & Stewart, 2013, pp. 9–10).

An advantage of supervised learning methods is that they require researchers to develop coherent definitions of concepts for particular applications, which leads to clarity in conceptualization and measurement. Supervised learning methods are also considered relatively easy to validate because there are clear statistics available that summarize model performance. Examples of computer science applications of supervised learning methods to sentiment analysis include recent work by Prabowo and Thelwall (2009), Shi and Li (2011), and Ortigosa-Hernandez and his colleagues (2012).

2. Metaphor Analysis

Metaphorical language provides striking evidence of how language is embodied (Lakoff & Johnson, 1980). Metaphors are highly parsimonious form-meaning

pairings because they express meaning by enacting bodily and emotional operations. Thus the formal analysis of metaphorical language is a second method for integrating information about bodily operations and capacities into automated semantic text analysis methods. Ignatow (2003, 2004, 2007) and several others (Schuster, Beune, & Stronks, 2011; Schmidt, 2012) have developed methods of metaphor analysis for use in social science research, but have thus far stopped short of developing highly automated methods that take advantage of currently available computing power, software and data. Today several research teams in computational linguistics and related fields are developing automated methods for automatically detecting metaphors in texts. Neuman and his colleagues (2013) have developed a number of interrelated algorithms that have proven highly accurate in identifying figurative versus non-figurative language (see also Gandy et al., 2013). The success of these projects points to the potential for automated methods of metaphor analysis to be used in social science applications in the near future.

3. Dictionary-Based Sentiment Analysis

The third method for integrating bodily operations and capacities into automated semantic text analysis involves use of dictionary-based sentiment analysis methods. Dictionary-based methods do not rely on ad hoc human coding of texts, but instead use off-the-shelf, validated sentiment lexicons. Dictionary methods are perhaps the most intuitive and easy to apply automated sentiment analysis method. They involve using the rate at which key words appear in a text to classify documents into categories or to measure the extent to which documents belong to particular categories. For example, dictionaries have been used to measure the tone of newspaper articles (Eshbaugh-Soha, 2010). Dictionary methods measure a document's tone by using a list of words with attached tone scores and the relative rate at which these words occur. A dictionary to measure tone is simply a list of words that are classified as positive or negative, either dichotomously or continuously. A number of off-the-shelf dictionaries are available that provide sentiment keywords (see Liu, 2012). Recent social science applications of dictionary methods include work by Young and Soroka (2011) and Eshbaugh-Soha (2010). Although use of off-the-shelf dictionaries is convenient, compiling new dictionaries is time-intensive. But compilation of such dictionaries may be necessary in many cases, as validation tests often find relatively poor accuracy when results from automated dictionary-based methods using off-the-shelf dictionaries are compared with results from human coders (see Duric & Song, 2012).

VI. CONCLUSIONS

All of the text analysis methods reviewed above require tradeoffs on the part of the researcher. While none are ideal, they are developing quickly, and have been

epistemically successful in fields such as linguistics and marketing. Implicit in all of these methods is the basic realist constructionist philosophical position that discourses are ontologically real entities that can be studied more or less like any other empirical phenomenon (Elder-Vass, 2012; Kaidesoja, 2013). Theorists and empirical researchers continue to debate what precisely discourses and culture actually are (e.g Archer, 1996; Bloch, 1998; Elder-Vass, 2012; Lizardo, 2007; Sperber, 1975; Turner, 2002), but the position of realist constructionists is clearly that they are real entities that can be investigated using a wide variety of empirical research methods including machine-assisted methods.

Given the ubiquity of digital information communication technologies, it should come as no surprise that social researchers are interested in finding ways to use accessible digital text archives, software platforms, programming languages, and social media resources for research purposes. Bail (2014) has recently surveyed some of the digital resources that are available for cultural research, and has implored sociologists to take maximum advantage of them. In a sense the purpose of the present paper has been to advance Bail's arguments by explicating metatheoretical and theoretical frameworks that can empower social scientists to use digital resources for cultural research. At the very least, it is my hope that this paper will generate discussion regarding how cultural researchers might take advantage of the vast information resources available in the digital age. Such resources can be used not only to amplify human interpretation through sheer computing power, but to supplement human interpretation through applications of sociologically and linguistically sophisticated algorithms.

Gabe Ignatow
Sociology
University of North Texas
1155 Union Circle #31157
Denton
Texas
United States
ignatow@unt.edu

REFERENCES

Alexander, J. (1990a). Introduction. In J. C. Alexander (Ed.), *Durkheimian sociology: Cultural studies* (pp. 1–22). Cambridge, UK: Cambridge University Press.

Alexander, J. (1990b). Culture and political crisis: "Watergate" and Durkheimian sociology. In J. C. Alexander (Ed.), *Durkheimian sociology: Cultural studies* (pp. 187–224). Cambridge, UK: Cambridge University Press.

Archer, M. (1996). *Culture and agency: The place of culture in social theory*. Cambridge, UK: Cambridge University Press.

Bail, C. (2014). The cultural environment: measuring culture with big data. *Theory and Society*, *43*(3–4), 465–482.

Bail, C. (2012). The fringe effect: civil society organizations and the evolution of media discourse about Islam since the September 11th attacks. *American Sociological Review*, *77*(7), 855–879.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609.

Bauer, M. W., Bicquelet, A., & Suerdem, A. K. (2014). Text analysis: An introductory manifesto. In M. W. Bauer, A. Bicquelet & K. S. Ahmet (Eds.), *Textual analysis*. London, UK: Sage.

Bergen, B., & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In J.-O. Östman & M. Fried (Eds.), *Construction grammars: Cognitive grounding and theoretical extensions* (pp. 147–190). Amsterdam: J Benjamins.

Biernacki, R. (2009). After quantitative cultural sociology: Interpretive science as a calling. In I. Reed & J. C. Alexander (Eds.), *Meaning and method: The cultural approach to sociology* (pp. 119–207). Boulder, CO: Paradigm Publishers.

Biernacki, R. (2012). *Reinventing evidence in social inquiry: Decoding facts and variables*. London: Palgrave Macmillan.

Bloch, M. (1998). *How we think they think: Anthropological approaches to cognition, memory, and literacy*. Boulder, CO: Westview Press.

Caren, N., Jowers, K., & Gaby, S. (2012). A social movement online community: Stormfront and the white nationalist movement. *Research in Social Movements, Conflict and Change*, *33*, 163–193.

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

Carley, K. (1997). Network text analysis: The network position of concepts. In R. Carl (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah, NJ: Lawrence Erlbaum.

Carley, K., & Palmquist, M. (1992). Extracting, representing and analyzing mental models. *Social Forces*, *70*(3), 601–636.

Carley, K., & Kaufer, D. (1993). Semantic connectivity: An approach for analyzing semantic networks. *Communication Theory*, *3*(3), 183–213.

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

Clark, A. (1997). *Being there: Putting brain, body, and world together*. Cambridge, MA: MIT Press.

Dreyfus, H. (2006). Overcoming the myth of the mental. *Topoi*, *25*(1–2), 43–49.

Duric, A., & Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. *Decision Support Systems*, *53*(4), 704–711.

Elder-Vass, D. (2010). *The causal power of social structures*. Cambridge, UK: Cambridge University Press.

Elder-Vass, D. (2012). *The reality of social construction*. Cambridge, UK: Cambridge University Press.

Elder-Vass, D. (2013). Debate: Seven ways to be a realist about language. *Journal for the Theory of Social Behaviour*, *44*(3), 249–267.

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

Fairclough, N., Jessop, B., & Sayer, A. (2007). Critical realism and semiosis. *Journal of Critical Realism*, *5*(1), 2–10.

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

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language*, *64*(3), 501–538.

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

Franzosi, R. (2010). Sociology, narrative, and the quality versus quantity debate (Goethe versus Newton), Can computer-assisted story grammars help us understand the rise of Italian fascism (1919–1922)? *Theory and Society*, *39*(6), 593–629.

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

Gandy, L., Allan, N., Atallah, M., Frieder, O., Howard, N., Kanareykin, S., Argamon, S. (2013). Automatic identification of conceptual metaphors with limited knowledge. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6398

Goldberg, A. (1995). *Constructions*. Chicago: University of Chicago Press.

Gottschalk, L. A., & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. Berkeley, CA: University of California Press.

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

Ignatow, G. (2003). Idea hamsters on the bleeding edge: profane metaphors in high technology jargon. *Poetics*, *31*(1), 1–22.

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

Ignatow, G. (2007). Theories of embodied knowledge: New directions for cultural and cognitive sociology? *Journal for the Theory of Social Behaviour*, *37*(2), 115–135.

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

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

Jacobs, R. N. (2000). *Race, media, and the crisis of civil society: From Watts to Rodney King*. Cambridge, UK: Cambridge University Press.

Kaufman, J. (2004). Endogenous explanation in the sociology of culture. *Annual Review of Sociology*, *30*(1), 335–357.

Kaidesoja, T. (2013). *Naturalizing critical realist social ontology*. New York: Routledge.

Lakoff, G. (2012). Explaining embodied cognition results. *Topics in Cognitive Science*, *4*(4), 773–785.

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

Langacker, R. W. (1999). *Grammar and conceptualization*. New York: Mouton de Gruyter.

Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 68–72.

Liu, B. (2012). *Sentiment analysis and opinion mining*. San Rafael, CA: Morgan & Claypool.

Lizardo, O. (2007). Mirror neurons, Collective objects and the problem of transmission: Reconsidering Stephen Turner's critique of practice theory. *Journal for the Theory of Social Behaviour*, *37*(3), 319–350.

Lizardo, O. (2009). Is a "special psychology" of practices possible? From values and attitudes to embodied dispositions. *Theory and Psychology*, *19*(6), 713–727.

Mast, J. L. (2006). The cultural pragmatics of event-ness: the Clinton/Lewinsky affair. In J. C. Alexander, B. Giesen & J. L. Mast (Eds.), *Social performance. symbolic action, cultural pragmatics, and ritual* (pp. 115–145). Cambridge: Cambridge University Press.

McLennan, G. (2005). The new American cultural sociology. *Theory, Culture and Society*, *22*(6), 1–18.

Mohr, J. W. (1998). Measuring meaning structures. *Annual Review of Sociology*, *24*, 345–370.

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

Mohr, J. W., & Rawlings, C. (2010). Formal models of culture. In J. Hall, L. Grindstaff & M. -C. Lo (Eds.), *A handbook of cultural sociology* (pp. 118–128). London: Routledge.

Mohr, J. W., & Rawlings, C. (2012). Four ways to measure culture: Social science, hermeneutics, and the cultural turn. In J. Alexander, R. Jacobs & P. Smith (Eds.), *The Oxford handbook of cultural sociology* (pp. 70–113). Oxford, UK: Oxford University Press.

Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Howard, N., & Frieder, O. (2013). Metaphor identification in large texts corpora. *PLoS ONE*, *8*(4), http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0062343

Ortigosa-Hernandez, J., Rodríguez, J. D., Alzate, L., Lucania, M., Inza, I., & Lozano, J. A. (2012). Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing*, *92*(1), 98–115.

Östman, J. -O., & Fried, M. (2004). *Construction grammars: Cognitive grounding and theoretical extensions*. Amsterdam: J Benjamins.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Infometrics*, *3*(2), 143–157.

Pugh, A. J. (2009). *Longing and belonging: Parents, children and consumer culture*. Berkeley, CA: University of California Press.

Reed, I. A. (2008). Justifying sociological knowledge: From realism to interpretation. *Sociological Theory*, *26*(2), 101–129.

Reed, I. A. (2011). *Interpretation and social knowledge*. Chicago: University of Chicago Press.

Reed, I. A., & Alexander, J. C. (2009). Social science as reading and performance: A cultural-sociological understanding of epistemology. *European Journal of Social Theory*, *12*(1), 21–41.

Roberts, C. W. (1997). Introduction. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. New York: Routledge.

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

Schmidt, R. (2012). Methoden der sozialwissenschaftlichen Metaphernforschung. *Metaphern und Gesellschaft*, *1*, 167–184.

Schrodt, P., & Savaiano, S. (1997). Environmental change and conflict: Analyzing the Ethiopian famine of 1984–85. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 147–158). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Sealey, A. (2009). Probabilities and surprises: A realist approach to identifying linguistic and social patterns, with reference to an oral history corpus. *Applied Linguistics*, *31*(2), 215–235.

Shalin, D. N. (2007). Signing in the flesh: Notes on pragmatist hermeneutics. *Sociological Theory*, *25*(3), 193–224.

Shi, H. -X., & Li, X. -J. (2011). A sentiment analysis model for hotel reviews based on supervised learning. Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, 10–13 July.

Sperber, D. (1975). *Rethinking symbolism*. Cambridge, UK: Cambridge University Press.

Sudhahar, S., Franzosi, R., & Cristianini, N. (2011). Automating quantitative narrative analysis of news data. *JMLR: Workshop and Conference Proceedings*, *17*, 63–71.

Sun, R. (Ed.). (2012). *Grounding social sciences in cognitive sciences*. Cambridge, MA: MIT Press.

Turner, M. (2001). *Cognitive dimensions of social science*. Oxford, UK: Oxford University Press.
Turner, S. P. (2002). *Brains, practices, relativism: Social theory after cognitive science*. Chicago: University of Chicago Press.
Young, L., & Soroka, S. (2011). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205–231.
Zerubavel, E. (2009). *Social mindscapes: An invitation to cognitive sociology*. Cambridge, MA: Harvard University Press.