Special Issue Article

Sociological Methods & Research
2022, Vol. 51(4) 1580–1633

DOI: 10.1177/00491241221099551
journals.sagepub.com/home/smr

Who Does What to Whom? Making Text Parsers Work for Sociological Inquiry

Oscar Stuhler[1]

Abstract

Over the past decade, sociologists have become increasingly interested in the formal study of semantic relations within text. Most contemporary studies focus either on mapping concept co-occurrences or on measuring semantic associations via word embeddings. Although conducive to many research goals, these approaches share an important limitation: they abstract away what one can call the event structure of texts, that is, the narrative action that takes place in them. I aim to overcome this limitation by introducing a new framework for extracting semantically rich relations from text that involves three components. First, a semantic grammar structured around textual entities that distinguishes six motif classes: actions of an entity, treatments of an entity, agents acting upon an entity, patients acted upon by an entity, characterizations of an entity, and possessions of an entity; second, a comprehensive set of mapping rules, which make it possible to recover motifs from predictions of dependency parsers; third, an R package that allows researchers to extract motifs from their own texts. The framework is demonstrated in empirical analyses on gendered interaction in novels and constructions of collective identity by U.S. presidential candidates.

[1]Department of Sociology, New York University, New York, NY, USA

Corresponding Author:
Oscar Stuhler, Department of Sociology, New York University, 295 Lafayette Street, 4th Floor, New York,
NY 10012-9605.
Email: stuhler@nyu.edu

Keywords

content analysis, text analysis, dependency parsing, computational sociology, semantic grammar, narrative, computational social science, word embeddings, cultural sociology, natural language processing

1. Introduction

Over the past decade, sociologists have become increasingly interested in the formal study of semantic relations within text. Encouraged by the increasing availability of large text corpora and computational resources, this research program comprises a variety of approaches. In what one might call semantic network analysis, researchers derive and study concept co-occurrence graphs (e.g., Hoffman et al. 2018; Fuhse et al. 2020; Lee and Martin 2014; Rule, Cointet and Bearman 2015; Padgett et al. 2020; Light 2014). Other studies aggregate co-occurrence structures to identify semantic clusters of frequently co-occurring terms – usually interpreted as topics (e.g., Fligstein, Stuart Brundage and Schultz 2017; DiMaggio, Nag and Blei 2013; Karell and Freedman 2019). More recently, scholars have moved towards using (neural) word embeddings to study distributions of semantic similarities (e.g., Kozlowski, Taddy and Evans 2019; Nelson 2021; Stoltz and Taylor 2021).

Despite the merits and promise of this burgeoning research program, most approaches currently used by sociologists share an important limitation: the relations under investigation are either ones of co-occurrence or, in the case of word embeddings, semantic distances derived from co-occurrence. These representational forms – although conducive to many research goals – are semantically underspecified in a way that limits our capacity to analyze texts. Specifically, they tend to abstract away what one can call the *event* structure of texts (Franzosi 1989:276; 2009:16–17), that is, the narrative action that takes place in a text. Many of the questions sociologists have traditionally had when studying texts, however, revolve around narrative events and action. Put simply, rather than whether A and B co-occur or are similar, sociologists often want to know what A *does* to B and vice versa. This becomes especially evident in work on accounts and narrative (Polletta et al. 2011; Franzosi 1998a; Abell 2004) but also in the discipline's pioneering works that formally studied relations in texts but extracted them via hand-coding (Carley 1993, 1994; Carley and Palmquist 1992; Bearman and Stovel 2000; Bearman, Faris and Moody 1999; Tilly 1997; Franzosi 1998b, 1994, 1989, 1990, Duquenne, Mohr and Le Pape 1998; Mohr 1994; Martin 2000; Smith 2007).

In this paper, I demonstrate that one way to overcome this limitation of most contemporary methods is to turn to dependency parsers – models that predict the syntactic relations between words within sentences. Dependency parsers hold great promise for sociological inquiry because they have the potential to extract semantically rich relations from textual data. However, only few sociologists have made use of dependency parsers in their research (recent examples are Goldenstein and Poschmann 2019; Stuhler 2021). This is surprising, given that automatic parsers have existed since at least the early 90s, and models that are deployable out-of-the-box while achieving high levels of accuracy have recently become available in software environments popular among sociologists, like R and Python.

I suspect that the main reason for this is a mismatch between the format of information that dependency parsers predict, and the kind of information sociologists want. Dependency parsers output fairly complex, directed graph structures in which edges between lexical units are labeled according to their *syntactic* dependency relation. Modern dependency grammars like the Universal Dependencies framework (Universal-Dependencies 2020) encompass more than 60 syntactic relation types. Sociologists, on the other hand, usually seek comparatively less complex *semantic* information – like the question of what A did to B, or what characteristics are attributed to them.

To overcome this mismatch, I propose a framework that involves three components. First, a simple, entity-centered semantic grammar that distinguishes six classes of motifs: *actions* of an entity, *treatments* of an entity, *agents* acting upon an entity, *patients*[1] acted upon by an entity, *characterizations* of an entity, and *possessions* of an entity. Second, a comprehensive set of inductively derived translation rules that map complex dependency trees into this grammar. Third, an R package that allows researchers to extract motifs from their own texts according to the proposed grammar.

I demonstrate the usefulness of this framework in two analyses, using text corpora from different discourses and of different sizes. First, I analyze gender relations in a corpus of U.S. Novels (1880–2000). I show that female-female interactions are similarly frequent as male-male interactions in novels written by women – but considerably less so in those written by men. I also investigate the asymmetry in inter-gender relations (e.g., do men kiss women or vice versa?) and trace its evolution over time. I find that gender has become less predictable based on interaction content over the 20th century. Second, I investigate U.S. presidential candidates' attempts

to construct a shared identity with their audience in campaign speeches (1950–2020). Specifically, I focus on the use of the first-person plural pronoun "we" and identify different rhetorics that are used to constitute this "we." Among other things, I show that in recent years there has been a tendency for candidates to increase their emphasis on campaign identity as Election Day approaches.

2. Relational Approaches to Text in Sociology

Most traditional quantitative content analyses proceed by designating particular key phrases or conceptual codes and then counting the presence or absence of these elements in a set of documents. In the late 1980s and 90s, however, several sociologists began to take a relational approach towards studying texts. Rather than assigning and counting codes, these approaches sought to formally analyze the *semantic relations* asserted by a text. Below, I review this research program, including its most recent developments. Because many of the approaches employed by sociologists have antecedents in other disciplines, it is worth emphasizing that my focus lies on work within sociology.

One of the earliest contributions to this program was Franzosi's proposal to study texts by disaggregating them into events of the form subject-action-object (SAO), constructs which he referred to as "semantic triplets" (Franzosi 1989, 1990, 1994). Similarly, Abell (1987) sought to define a general formalism for representing narrative structures. Also among the early contributions to this program are Roberts' (1989, 1997) efforts to design a generic semantic grammar. His formally ambitious grammar, however, did not only aim at parsing event clauses into their constituents, but also at classifying them according to their discursive function. Finally, Carley (1993) developed MAP analysis – an approach she saw in explicit contrast to traditional content analysis. The core innovation of MAP analysis was to represent textual data as networks in which concepts form nodes and edges specify the relationship between them. Which kinds of relationships are permissible and which concepts are relevant is left to the researcher (Carley and Palmquist 1992). In most applications, this approach was used to generate network representations of mental models from interview transcripts (Carley 1988; Carley and Palmquist 1992), but others include a study on shifting images of robots in science fiction novels (Carley 1994).

These initial contributions were followed by a set of studies that sought to apply this form of analysis to distinct problem areas. Some of them focused on the differentiation among different kinds of identities. Notable examples include Mohr's contributions on categories of poverty

and their association with particular forms of aid (Mohr 1994; Mohr and Duquenne 1997; also see Mohr and Lee 2000) and Martin's (2000) work on animals and their occupations in children's literature. Yet another group of scholars focused on studying the structures of narrative networks, in which nodes are events and edges are causal or logical links between them asserted by the text (Bearman and Stovel 2000; Bearman et al. 1999; Smith 2007). Meanwhile, extracting relations from textual data also became popular among social movement and contentious politics scholars. In one of the most impressive applications of a semantic triplet grammar, Tilly (1995, 1997) analyzes contentious gatherings in 18th and 19th century Great Britain. From a corpus of newspapers, he extracted more than 50,000 reports of relational actions between classes of political actors. These were then subjected to a blockmodel analysis in order to identify different political factions. Similar semantic grammars have since been applied to code newspaper reports about protest events and political claims-making (e.g., Koopmans and Statham 1999; Wada 2004). Finally, another class of studies from this early phase of relational content analysis builds on the duality of texts and textual elements. Examples of this include Mische and Pattison's (2000) analysis of claims in manifestos and speeches of political organizations in Brazil, McLean's (1998) work on rhetorical elements in favor-seeking letters in Renaissance Florence, as well as Ruef's (1999) analysis of health related activities and organizational forms in a corpus of medial publications. These works are conceptually distinct from those above, however, in so far as the relations between textual entities arise from their distribution over texts, rather than from their meaningful association *within a text*.

In the past decade, sociologists have increasingly turned to automating the process of extracting semantic relations from text. Perhaps the most straightforward approach is what one may call word network analysis, which has some history outside of sociology (e.g., Danowski 1993). In this approach, text is transformed into a graph where nodes represent words or phrases and edges represent some measure of co-occurrence within higher units of text like documents, paragraphs, or moving windows. In some cases, a community detection algorithm is then applied to detect topical clusters within the network. Examples of this approach include Leydesdorff and Hellsten's (2006) analysis of scientific controversies, Light's (2014) study of United States' presidential inaugural, Lee and Martin's (2014; also see Lee and Martin 2014) mapping of different Frankfurt School thinkers' conceptual toolkits, Rule et al.'s (2015) analysis of changes in the U.S. State of the Union address, Fuhse and colleagues' (2020) investigation into the multiple

meanings of "Volk" in Germany's Weimar republic, and Padgett et al.'s (2020) analysis of the discourse in Florentine Consulte e Pratiche council.".

A related approach that has become widely popular among sociologists in recent years is topic modeling, where thematic structures are inferred from the co-occurrence of words in higher units of text (e.g., DiMaggio et al. 2013; Fligstein et al. 2017; Karell and Freedman 2019; Mohr and Bogdanov 2013). Topics are formalized as probability distributions over the feature vocabulary. Apart from its inductiveness, however, this approach bears a closer resemblance to traditional content analysis: rather than investigating the relations between textual entities, the main goal is usually to label documents or their subsections with regard to the relative prevalence of different topics.

An approach more immediately concerned with relations between textual units is the study of semantic similarity via word embeddings, also known as vector semantics. Here, the words in a corpus are represented as equal-length numerical vectors – they are *embedded* in a vector space. There are different ways to construct these vector spaces. In principle, one may consider the vector associated with a word in a document-term matrix to be an embedding. Usually, however, additional processing steps are undertaken to make the embeddings relatively short, dense, and informative regarding the semantic content of a word – which conceptually moves similarity of the embeddings towards notions of semantic similarity and away from the idea of co-occurrence. For instance, in their analysis of the Protestant Bible and its use, Hoffman and colleagues (2018) first generate a word-to-word matrix in which a cell captures the number of times two words co-occur in a moving text window (weighted by distance in the text), much like in the co-occurrence based approaches discussed above. Instead of studying this matrix itself as a network, however, the authors treat it as vector space and compute the cosine similarities between each word's embedding vector. The authors then move on to generate and study a graph in which nodes are words and edges are semantic similarity measured as cosine similarities between embeddings (for a similar approach, see Puetz, Davis and Kinney 2021).

More elaborate techniques for generating word embeddings have existed for some time (e.g., Deerwester et al. 1990), but the breakthrough of this approach in sociology came with the ability to efficiently train high quality word embeddings via the widely popular *word2vec* (Mikolov et al. 2013) and *GloVe* (Pennington, Socher and Manning 2014) algorithms (for social science application oriented reviews, see Arseniev-Koehler 2021; Spirling and Rodriguez 2022). In an original application, Kozlowski and colleagues

(2019) use *word2vec* embeddings trained on the Google Ngram Corpus to generate measures for a number of cultural dimensions (e.g., male – female, rich – poor, moral – immoral). This is done by averaging the differences of the embeddings of multiple word pairs representing a dimension (e.g., "man" and "woman", "male" and "female", "him" and "her"). The so-generated vectors are then used to trace the association between different cultural dimensions over time and to project words representing cultural practices onto them. With slight variations, this approach has been adopted in a number of very recent sociology papers. Examples include studies on gender and its changing associations with educational stereotypes (Boutyline, Arseniev-Koehler and Cornell 2020) or social domains broadly (Jones et al. 2020), scientific concepts and their differential evaluation in quantitative and qualitative research cultures (Kang and Evans 2020), obesity-related concepts and their associations with different cultural dimensions (Arseniev-Koehler and Foster 2020), the evolution of the U.S. immigration discourse (Stoltz and Taylor 2021), the intersectionality of race and gender in the nineteenth-century U.S. South (Nelson 2021), and changes in the semantic environment of the "network" concept (Yung 2021). Slightly different is the approach developed by Stoltz and Taylor, which centers around estimating the semantic similarity of whole documents with focal concepts (2019) or cultural dimensions (Taylor and Stoltz 2021). In a recent extension, the authors show how this approach can be used to measure schemata underlying a text (Taylor and Stoltz 2020).

Table 1 gives an overview of sociological work that formally engages with semantic relations in textual data. It lists the kinds of textual concepts studied and the types of relationships between them, as well as how these elements were extracted. Earlier works that relied on hand-coding overwhelmingly had a defined interest in particular relations or classes thereof: identity categories being *treated* (Mohr), political actors *acting* towards other political actors (Tilly), animals *doing* jobs (Martin), events being *caused* by other events (Bearman and colleagues) and actors, or, more broadly, subjects *acting* towards objects (Franzosi, Abell, Carly). In contemporary work, there continues to be variation in subject matters, but with the turn towards computational approaches the literature seems to have largely converged on two kinds of relationships: co-occurrence and semantic similarity.

What should we make of this development? Unquestionably, recent methodological advances have led the formal, relational study of text to a new level of popularity among sociologists. What was once a niche discourse can now be found in the discipline's flagship journals. Nonetheless, the early studies that went through the trouble of manually extracting semantic

Table 1. Overview of Formal, Relational Approaches to Text in Sociology.

Publications	Extraction methods	Classes of textual concepts	Relation types
(Abell 1987)	Hand coding	[Agents] [Events]	[Bringing about]
(Carley 1988, 1993, 1994; Carley and Palmquist 1992)	Computer assisted hand coding	[Concepts]	[Relationship] Note: Carley leaves it up to the researcher to define classes of eligible relationships.
(Franzosi 1989, 1990, 1994)	Hand coding	[Subjects] [Objects]	[Actions]
(Roberts 1989, 1997)	Hand coding	[Subjects] [Objects] [Actions] [Discusive function] [Mode] [Tense]	Probability Note: Roberts' full scheme encodes clauses into 17 (1997) distinct variables and then investigates statistical associations between them.
(Mohr 1994; Mohr and Duquenne 1997),	Computer assisted hand coding	[Categories of poverty] [Forms of aid]	Treatment
(Mohr and Lee 2000)	Computer assisted hand coding	[Identity discourses] [Outreach practices]	Treatment
(Tilly 1995, 1997)	Hand coding	[Political actor]	[Directed political actions]

(continued)

Table 1. Continued

Publications	Extraction methods	Classes of textual concepts	Relation types
(McLean 1998)	Hand coding	[Keywords]	Co-occurrence
(Ruef 1999)	Computational	[Organizational forms] [Health related activities]	Co-occurrence
(Bearman and Stovel 2000; Bearman et al. 1999; Smith 2007)	Hand coding	[Events]	Causal or logical connection
(Martin 2000)	Hand coding	[Animals] [Jobs]	Doing
(Mische and Pattison 2000)	Hand coding	[Political goals]	Co-occurrence
(Light 2014; Leydesdorff and Hellsten 2006; Fuhse et al. 2020; Lee and Martin 2014; Padgett et al. 2020)	Computational	[Words]	Co-occurrence
(DiMaggio et al. 2013; Fligstein et al. 2017; Karell and Freedman 2019; Mohr and Bogdanov 2013)	Computational	[Topics] [Words]	Probability
(Mohr et al. 2013)	Computational	[Concepts]	[Actions]
(Rule et al. 2015)	Computational	[Noun phrases]	Co-occurrence
(Lee and Martin 2018)	Computational	[Author mentions]	Co-occurrence

(continued)

Table 1. Continued

Publications	Extraction methods	Classes of textual concepts	Relation types
(Hoffman et al. 2018)	Computational	[Words]	Semantic similarity
(Kozlowski et al. 2019)	Computational	[Cultural dimensions]	Semantic similarity
(Stoltz and Taylor 2019)	Computational	[Focal concepts] [Documents]	Semantic similarity
(Boutyline et al. 2020)	Computational	[Educational stereotypes] Gender dimension	Semantic similarity
(Jones et al. 2020)	Computational	[Social domains] Gender dimension	Semantic similarity
(Arseniev-Koehler and Foster 2020)	Computational	[Cultural dimensions] [Obesity related words]	Semantic similarity
(Kang and Evans 2020)	Computational	[Words] Evaluative dimension	Semantic similarity
(Taylor and Stoltz 2021; Taylor and Stoltz 2020)	Computational	[Cultural dimensions] [Documents]	Semantic similarity
(Nelson 2021)	Computational	[Cultural dimensions] [Social institutions]	Semantic similarity
(Stoltz and Taylor 2021)	Computational	[Cultural dimensions]	Semantic similarity

(continued)

Table 1. Continued

Publications	Extraction methods	Classes of textual concepts	Relation types
		[Immigration related words]	
(Yung 2021)	Computational	[Words] Network concept	Semantic similarity
(Puetz et al. 2021)	Computational	[Noun phrases]	Semantic similarity

Note: Rows are ordered by the respectively first publication year. Square brackets denote sets in the sense that not all elements are identical. This table is the author's best effort to provide a coarse overview of a large number of different approaches and empirical studies. Some authors might rightfully find the complexity of their approach to be underrepresented (e.g., Roberts 1989, 1997). Others define their schemes in explicitly flexible terms (e.g., Abell 1989) or run varied analyses (e.g., Kozlowski et al. 2019, Stoltz and Taylor 2021), so that their representation in the table is according to what appears to be the main application.

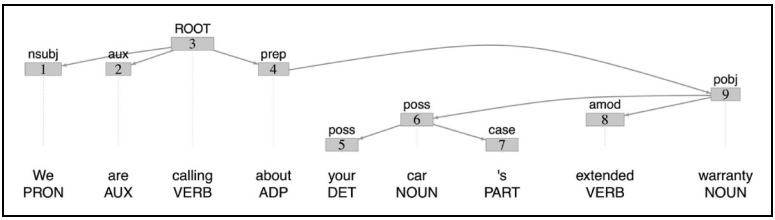

Figure 1. A sentence and corresponding dependency tree.
Note: Syntactic relationships between tokens are directed as indicated by the arrows pointing from the head to the dependent token. The depiction of dependency trees in this paper follows the convention of placing the tag of a dependency relation above the dependent node. For instance, "We" is the nominal subject (nsubj) of "calling.".

relations from texts are a testament to the fact that many sociological research questions go beyond co-occurrence and semantic similarity. They concern *narrative events* (Franzosi 1989:276; 2009:16–17) in which textual entities engage in actions, or are recipients of them. Notably, this implies that they revolve around kinds of relations that the currently dominant approaches do not capture. In what follows, I describe a way to overcome this limitation and thereby attempt to reconcile the contemporary literature with its origins.

3. What are Dependency Parsers?

One way to overcome the described limitations of many contemporary methods is to turn to dependency parsers. Dependency parsers hold great promise for sociological inquiry because they have the potential to extract semantically rich relations from textual data. However, within sociology, only few have made use of this opportunity (Stuhler 2021; Goldenstein and Poschmann 2019; Mohr et al. 2013; in political science, see van Atteveldt, Kleinnijenhuis and Ruigrok 2008; van Atteveldt et al. 2017). My ambition here is first, to identify and explicate some of the intricacies involved in making dependency parsers work for sociological inquiry; second, to provide a framework and associated software that overcomes these problems; and third, to showcase the potential of this framework in empirical analyses. I begin with a review of what dependency parsers do (for more extensive introductions, see Jurafsky and Martin 2020; Eisenstein 2019:243–266).

Dependency parsers are models that predict the *syntactic* relations between lexical items (mostly words) within sentences. Figure 1 shows an English sentence that is annotated with a *dependency tree*. Dependency trees are directed graphs that specify relationships between the lexical items of a sentence. What counts as a permissible dependency tree depends on the *dependency grammar* (a glossary of potentially unfamiliar terms is provided in the supplementary material). Most grammars share a series of constraining properties: there is one *root* node with in-degree 0 (usually the main verb of the sentence); all other nodes have an in-degree of 1, or, put differently, each non-root node is a *dependent* to exactly one *head* node; the graph is fully connected; there are no cycles in the graph. The justification and origin of these

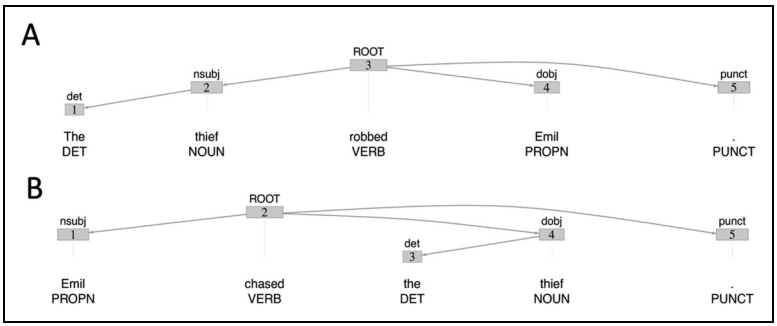

Figure 2. Two simple sentences with corresponding dependency tree.

constraining properties lie beyond the scope of this paper (for a good review, see de Marneffe and Nivre 2019).

Besides these structural properties of dependency trees, dependency grammars define the range of possible dependency types. These specify the grammatical function that a dependent plays with regard to its head. Throughout this paper, I build on the ClearNLP scheme (Choi and Palmer 2012), which is closely related to the Stanford dependency scheme (de Marneffe and Manning 2008). In Figure 1, for instance, "We" is the nominal subject (nsubj) of the verb "calling" and the edge between the two terms is accordingly labeled. The root of the sentence, "calling," has two more dependents: "are" which serves as an auxiliary (aux) and "about" which introduces a prepositional phrase (prep).

Dependency parsers are trained on and evaluated against datasets with annotated syntactic relations called treebanks. For English, most of these datasets are generated by applying transformation rules to existing constituency treebanks (e.g., Choi and Palmer 2012). The models that were trained on these data are then made available as software and can be deployed more or less out of the box on other texts. In practice, dependency parsers are usually embedded in comprehensive annotation pipelines that process raw textual data by sequentially solving a set of tasks, including word tokenization, sentence splitting, part-of-speech tagging, and lemmatization. The information from these preceding tasks is the basis for parsing (for introductions on the mechanics of dependency parsers, see Jurafsky and Martin 2020; Eisenstein 2019:243–266). There are multiple popular processing pipelines which include dependency parsing (e.g., Stanford CoreNLP by Manning et al. 2014; or UDPipe by Straka 2018). For this paper, I use a pipeline[2] provided by the spaCy library (Honnibal et al. 2020), which has recently been implemented in R (Benoit and Matsuo 2020).

Dependency trees provide *syntactic* relationships – that is, relationships regarding the grammatical structure of a sentence. Syntactic relationships may not be of interest *per se*, to most sociologists. However, they can provide an approximation of the *semantic* relationships – that is, relationships regarding the meaning of a sentence. To illustrate this, consider the sentence in Panels A and B of Figure 2, and imagine we are interested in actions "Emil" is involved in. An analysis of co-occurrences would allow us to infer that Emil is associated with "robbing" and with "chasing," but whether he is chased and robbed or chasing and robbing or a combination of these two remains unclear. By building on the syntactic relationships, we can go beyond this information and infer whether he is the agent (doer)

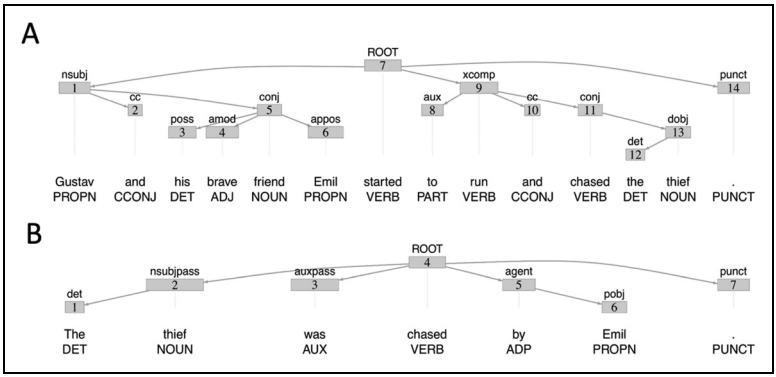

Figure 3. Two complex sentences with corresponding dependency tree.

or patient (recipient) of these actions. To do this, we could write two simple rules:

(a) Find all verbs "Emil" is the nominal subject of, that is, all nsubj-heads of "Emil."
(b) Find all verbs "Emil" is the direct accusative object of, that is, all dobj-heads of "Emil."

Applying these rules, we could now infer that Emil is robbed and chasing. If we wanted to go beyond this, we might expand our rules to include any dobj-dependents of Emil's nsubj-heads, as well as the nsubj-dependents of Emil's dobj-heads. This would allow us to infer that Emil was robbed by a "thief" and is chasing a "thief"; thereby we would capture quite literally the entire meaning of both sentences.

While we would get beyond simple co-occurrence with our two rules, there is a problem: most sentences we encounter in real textual data are syntactically more complex than the example. For reference, the average sentence in the New York Times contains 23 words (Tauberg 2019). To illustrate this, consider the sentence in Panel A of Figure 3. This sentence has quite a bit more going on than the first example. Among other things, we learn that Emil is characterized as "brave" and that he is "running." Nonetheless, a *semantic* reading of this still relatively short sentence informs us that Emil is chasing a thief. However, things get considerably more complicated when we turn to the *syntactic* relations of the sentence.

First, "Emil" is an appositional modifier (appos) of the term "friend"; "friend" is linked to "Gustav" via a conjunct relationship (conj); "Gustav" is the nominal subject (nsubj) of "started" – the root of the sentence; "started" is linked to the verb "run" via an open clausal complement relationship (xcomp); "run" is linked via another conjunct relationship (conj) to "chased"; finally, "chased" again has the direct object (dobj) "thief." Rather than a direct connection and one syntactic relation type, we now have a path length of five and five distinct syntactic relation types between "Emil" and "chase." The sentence in Panel B of Figure 3 illustrates another point: sentences need not necessarily add much complexity to throw us off with our two rules. The syntactic relations of a simple passive clause look quite different than those of an active one.

As Franzosi pointed out long ago, sociologists are usually concerned with semantic, rather than syntactic relationships (Franzosi 1989:271–272). The syntactic relations derivable via dependency parsing can provide a basis for establishing semantic ones. They are, however, too fine-grained to be of much immediate use to sociologists, who usually have little regard for categories such as "appositional modifier" or "open clausal complement." The few sociology papers that have used parsers have left this distinction largely unacknowledged (Goldenstein and Poschmann 2019; Mohr et al. 2013; for the same point of criticism, see Monroe 2019; though see Stuhler 2021). Making dependency parsers work for sociological inquiry requires us to recognize and address this mismatch.

4. From Syntax to Semantics: Proposing an Entity-Centered, Semantic Grammar

How can we overcome the gap between the fine-grained syntactic relations predicted by dependency parsers and the semantic information most sociologists care for?[3] In this section, I propose a relatively simple, entity-centered semantic grammar that involves six classes of elements: *actions* of an entity, *treatments* of an entity, *agents* acting upon an entity, *patients* acted upon by an entity, *characterizations* of an entity, and *possessions* of an entity. I will make the case that we can map the complex syntactic relations predicted by dependency parsers into this semantic grammar via a set of transformation rules. Going forward, I refer to elements of the semantic grammar as *motifs*. I use semantic categories like agent, action, and patient rather than syntactic categories like subject, predicate, and object to highlight that motifs are semantic elements.

The grammar proposed here is based on what Franzosi called a semantic triplet, which, in slight variations, has been a canonical core of semantic grammars for textual microstructures (see e.g., Franzosi 1989:273–274; Propp [1928] 1968:113; van Dijk 1972:287; Todorov and Weinstein 1969:74). The main difference is that my grammar is entity-centered, so that motif classes are distinguished with regard to the relationship that they take towards a *core entity* of interest. This choice anticipates a use case in which researchers are interested in extracting semantic relationships around a particular concept or class thereof. Rather than to represent whole texts, the goal is to extract and represent the claims a text makes regarding these

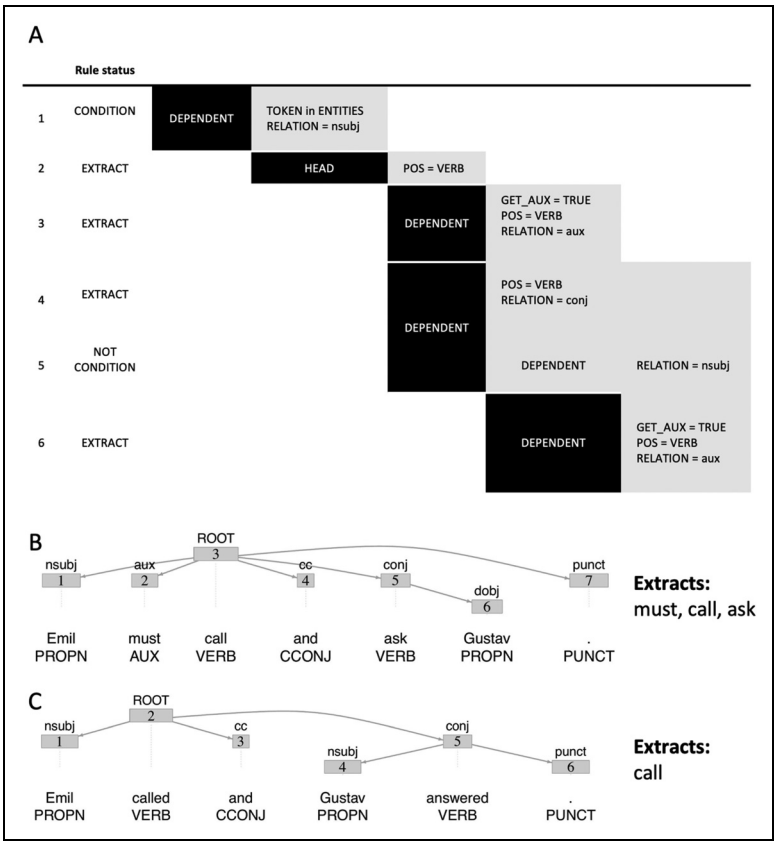

Figure 4. Extraction rule and two exemplary sentences.

entities. The work reviewed above corroborates that this is a central concern of research. Furthermore, by including *characterizations* and *possessions*, I add two stasis motifs – motifs that relate to states rather than to happenings (for a discussion of stasis elements in narratives see Franzosi 2009:18–19; Chatman 1978:31–33).

The mapping of the syntactic structures predicted by the dependency parser into the proposed semantic grammar occurs via a set of extraction rules. These rules were created in an iterative, inductive process. First, I started out with a set of simple extraction rules like the two formulated in the Section 3. Then, I applied these rules to a set of texts to extract motifs. I sampled sentences and, based on a semantic reading of the text, evaluated whether the rules would leave out textual elements that should have been extracted as motifs or extract ones that should not have been. In this respect, the previous section illustrated how a simple rule (extract all nsubj-heads of an entity) succeeds in extracting all *actions* in one context (Sentence B in Figure 2) but fails to do so in others (Sentences A and B in Figure 3). These evaluations led, with decreasing probability as the process went on, to either additions of new rules or refinements of existing rules. The rules specify criteria with regard to tokens, part-of-speech tags, and dependency relations. Annotated code implementing these rules is made available in the form of an R package, *semgram*, that is currently in development and hosted on Github.[4] *semgram* builds on functionalities of *spacyr* (Benoit and Matsuo 2020) for dependency parsing and *rsyntax* (Welbers, van Atteveldt and Kleinnijenhuis 2021) for implementing extraction rules.

While the rules themselves are intricate and too numerous to be discussed here, their general logic is demonstrated with an example. Consider the rule showcased in panel A of Figure 4, which extracts action motifs from the text around a core entity of interest. Imagine we apply this rule for extracting actions of Emil (hyperparameter: ENTITIES = ["Emil"]) in the sentences in panels B and C; we also specify that we are interested in extracting auxiliaries as actions (hyperparameter: GET_AUX = TRUE). The first line looks for all lexical items the token representation of which is among the entities of interest (TOKENS IN ENTITIES) that are nominal subject dependents (RELATION = nsubj) of a head. Line 2 proceeds to extract all heads that are verbs (POS = VERB). This leads to the extraction of "call" in both sentences. In line 3, we proceed to extract all verbial (POS = VERB) auxiliary (RELATION = aux) dependents of the elements identified in line 2. This leads to the extraction of "must" in the sentence in Panel B. Note that we do this because we specified to extract auxiliaries as actions (GET_AUX = TRUE), which may be undesirable for some applications. In line 4, we also

Table 2. Exemplary Sentences and Extractions for Different Motif Classes.

Motif class	Ex.	Sentence	Extract	Comment
Action	a.1	ENTITY calls.	*a_call*	Verbial heads ("calls") of which ENTITY is the nominal subject are the most frequent *actions*.
	a.2	ENTITY can call.	*a_can, a_call*	Auxiliary verbs can be considered *actions* (hyperparameter).
	a.3	ENTITY calls and asks.	*a_call, a_ask*	Conjunct verbs are considered *actions*.
	a.4	John and ENTITY called.	*a_call*	Instances in which ENTITY is a conjunct dependent of the nominal subject ("John") are considered for identifying *actions*.
	a.5	John was called by ENTITY.	*a_call*	Passive constructions with "by" are considered for identifying *actions*.
	a.6	My friend ENTITY called John.	*a_call*	Instances in which ENTITY serves as appositional modifier of a nominal subject (friend) can be considered *actions* (hyperparameter).
	a.7	ENTITY wants to call.	*a_want, a_call*	A verb in a clausal component is considered an *action*, as long as it doesn't have a nominal subject dependent.
	a.8	ENTITY wants you to call.	*a_want*	
Patient	P.1	ENTITY asks John.	*P_John*	*Patients* are usually direct objects of all transitive verbs identified as actions.
	P.2	Peter and ENTITY ask John.	*P_John*	
	P.3	My friend ENTITY asks John.	*P_John*	
	P.4	ENTITY came and asked John.	*P_John*	
	P.5	ENTITY wants to ask John.	*P_John*	
	P.6	ENTITY calls John, Jane, and Steve.	*P_John, P_Jane, P_Steve*	Multiple conjunct objects are simultaneously considered as *patients*.
	P.7	ENTITY asks	*P_John, P_question*	Dative objects ("John") of

(continued)

Table 2. Continued

Motif class	Ex.	Sentence	Extract	Comment
		John a question.		*actions* are also considered *patients*.
	P.8	John is asked by ENTITY.	*P_John*	Nominal passive subjects ("John") are also considered *patients*.
Treatment	t.1	John calls ENTITY.	*t_call*	*Treatments* are usually verbs of which ENTITY is the direct or dative object.
	t.2	John gives ENTITY an apple.	*t_give*	
	t.3	John gives Peter an ENTITY.	*t_give*	
	t.4	John calls Peter and ENTITY.	*t_call*	Instances in which ENTITY is a conjunct of a direct or dative object, are considered in extracting treatments.
	t.5	John gave Peter an apple and ENTITY.	*t_give*	
	t.6	ENTITY was called.	*t_call*	Verbs of which ENTITY is a nominal passive subject are treatments.
Agent	A.1	John calls ENTITY.	*A_John*	*Agents* are often the nominal subjects of verbs identified as *treatments* of ENTITY.
	A.2	John gives Peter an ENTITY.	*A_John*	
	A.3	Peter and John ask ENTITY.	*A_Peter, A_John*	Rules specified with regard to the relationship between *ENTITY* and *actions* apply to the relationship between *agent* and *treatment*, including conjunct subjects, conjunct verbs, appositions, or clausal components.
	A.4	ENTITY is asked by John.	*A_John*	
	A.5	John came and asked ENTITY.	*A_John*	
	A.6	John wants to ask ENTITY.	*A_John*	
	A.7	My friend John asked your brother ENTITY.	*A_friend, A_John*	

(continued)

extract verbial conjunction (RELATION = conj) dependents of the line 2 elements, which leads to the extraction of "ask" in the sentence in Panel B. By itself, this would also lead to the extraction of "answered" in the sentence in panel C, which Emil is clearly not the agent of. It was only through the

Table 2. Continued

Motif class	Ex.	Sentence	Extract	Comment
Characterization	be.1	ENTITY is kind.	*be_kind*	Most *characterizations* are adjectival dependents of a copula verb (be, become, remain, feel, look, and others).
	be.2	ENTITY looks sad.	*be_sad*	
	be.3	ENTITY is the winner.	*be_winner*	*Characterizations* can also be nominal attribute dependents of a copula verb.
	be.4	ENTITY remained president.	*be_president*	
	be.5	ENTITY could be the president.	*be_president*	Modifications via auxiliary verbs ("could") don't affect the *characterization* status.
	be.6	ENTITY is kind and honest.	*be_kind, be_honest*	Conjunct dependents of *characterizations* are considered *characterizations*.
	be.7	ENTITY won but remained sad.	*be_sad*	As with actions, the copular verb can take different positions for its adjectival or nominal dependent to be considered a *characterization*. Adjectival modifiers are considered *characterizations*.
	be.8	ENTITY is going to be sad.	*be_sad*	
	be.9	ENTITY hopes to remain president.	*be_president*	
	be.10	John bought a cheap, new ENTITY.	*be_cheap, be_new*	Adjectival modifiers are considered *characterizations*.
	be.11	The winner was ENTITY.	*be_winner*	Nominal subjects of a copular verb with ENTITY as attribute dependent can be considered *characterizations* (hyperparameter).
	be.12	The winners were John and ENTITY.	*be_winner*	
	be.13	My brother ENTITY won.	*be_brother*	When entity serves as appositional modifier, its head is considered a characterization.
Possession	H.1	ENTITY's spouse, friends, and parents were shocked.	*H_spouse, H_friend, H_parent*	Nouns, together with their conjunct dependents, that have ENTITY as possession modifier are considered *possessions*.
	H.2	The breaks and		Nominal heads of a preposition

(continued)

Table 2. Continued

Motif class	Ex.	Sentence	Extract	Comment
		wheels of the ENTITY were old.	*H_break, H_wheel*	with "of" that has ENTITY as object dependent are considered *possessions*.
	H.3	ENTITY has friends and enemies.	*H_friend, H_enemy*	Direct objects of the verb "have" and its inflections are considered *possessions*. Note: "Have" and its inflections is not considered as action motif. It's direct object or dative dependents are not considered patients.
Action-patient	aP.1	ENTITY asks John.	*aP_ask_John*	This is a composite motif class, taking all *patient* motifs and merging them with the respective *action*.
	aP.2	ENTITY made and ate a cake.	*aP_eat_cake*	A noun can be the object of multiple transitive verbs at the same time ("I made and ate the cake."), but the syntactic grammar used here does not allow to make inference regarding the relationship between a second verb and the object, so that there is only one action per patient in action-patient motifs (and only one treatment per agent in agent-treatment motifs).
Agent-treatment	At.1	John asks ENTITY.	*At_John_ask*	This is a composite motif class, taking all *agent* motifs and merging them with the respective *treatment*.

process of testing rules on actual data, that I realized that such a rule would lead to erroneous extractions whenever a second nominal subject ("Gustav" in this case) was a dependent of the conjunct verb. This led me to revise the rule and add the negative condition in line 5, specifying that line 4 elements would only be extracted if they did not have a nominal subject dependent (RELATION = nsubj). Finally, in line 6, we extract verbial auxiliary dependents of line 4 elements. This leads to no further extractions in our two sentences.

Table 2 gives examples for each motif class and lists some more specific cases to illustrate the scope of the rules. These cases may not seem complex in their semantics. Perhaps counterintuitively, however, the underlying syntactic structures are such that they require special consideration. The comment column of Table 2 and the discussion below provide further elaborations concerning this. Note, furthermore, that the table is not exhaustive of all possible scenarios and rules. In fact, the complexity of the formal rule set mostly arises from the necessity to account for the combination of different scenarios shown in Table 2 – say a combination of appositional modification, subject and verb conjuncts and an open clausal component, as seen in the sentence in Panel B of Figure 3. Table 2 also illustrates the markup-style used to distinguish the different motif classes which, as we will see below, can be useful for processing and analyzing the extracted motifs. Letters at the beginning of words indicate the respective motif class – so that a_call, for instance, implies that the entity of interest engaged in the *action* of calling, whereas t_call is used when calling is a *treatment* of the entity.

Action motifs imply that the entity of interest is doing something. The most straightforward example of this is when the entity serves as a nominal subject of a verb (example a.1). There are various syntactic constructions, however, in which a verb is considered an action despite the entity not being its nominal subject. This includes instances in which the entity is the conjunct of a nominal subject (a.4), there are multiple verbs (a.2, a.3, a.7), the entity serves as an appositional modifier of a nominal subject (a.6), and passive constructions (a.5). All actions are either lexical verbs or, if explicitly specified, auxiliary verbs.

Patient motifs are things that the entity of interest acts towards. They are usually objects of transitive verbs that were identified as an entity's action. These objects can be in accusative case (P.1-P.6) or in dative case if the verb is ditransitive (P.7). Any *action* motif can lead to multiple *Patient* motifs – as any transitive verb can have multiple conjunct objects (P.6). Beyond objects, nominal passive subjects are also considered patients (P.8).

Treatment motifs imply that something is done to an entity of interest. This is the case when the entity is the object of a transitive verb. The relationship between *treatments* and the entity is analogous to that of *actions* and *patients*. The entity can function as accusative (t.1, t.2) or dative (t.3) object, as nominal passive subject (t.6), or as conjunct of any of these (t.4, t.5).

Agent motifs are things that act towards the entity of interest via a *treatment* motif. In most cases, agents are the nominal subject of a verb that has been identified as a *treatment* motif (A.1, A.2). However, *agents* need not take that position and can be conjuncts (A.3) or appositional modifiers

(A.7) of the nominal subject. Generally, the relationship between *agents* and *treatments* is analogous to that of the entity and actions, so that the transitive verb may take different positions (A.5, A.6), and passive constructions in which the entity serves as nominal passive subject (A.4) are considered.

Beyond these process motifs, there are two classes of stasis motifs. *Characterizations* are characteristics ascribed to the entity of interest. There are several ways in which this can happen. The most common one is via a copular verb, that has either an adjectival (be.1, be.2, be.6, be.7, be.8) or nominal (be.3, be.4, be.5, be.9) attribute dependent. However, adjectives can also be direct dependents of the entity (be.10) to be considered *characterizations*. Furthermore, nominal subjects of copular verbs with the entity as attribute dependent (be.11, be.12) and heads with the entity as appositional modifier (be.13) are considered characterizations.

Possessions are things that the entity of interest is said to possess. The rule set accounts for three ways in which this can be expressed. First, when the entity serves as a possession modifier to a noun, said noun and its conjunct dependents are considered possessions (H.1). Second, constructions where the entity serves as object dependent of the preposition "of" can lead to possessions (H.2). Third, if the entity serves as nominal subject of "have" or one of its inflections, its direct object and nominal conjunctions thereof are considered *possessions*. Note that "have" is a transitive verb, but within the grammar, it is not considered an *action*, and consequently its objects aren't considered *patients*.

Finally, there are two composite motif classes which link *actions* and *patients* (aP.1), as well as *agents* and *treatments* (At.1). These are essentially what Franzosi called semantic triplets centered around the entity of interest. While representing data in this way can lead to high levels of sparsity as some have pointed out (Monroe 2019), I will demonstrate in section 5.1 that it can be a powerful means to analyzing representations of social relationships.

5. Application

Why should sociologists care about semantic grammars and the possibility to extract motifs from text? In this section, I use the introduced framework to conduct analyses concerning two classic sociological subjects: gender relations and collective identity. I will demonstrate how the framework provides new analytic purchase on these issues.

5.1 Who Kisses Whom? Gender Relations in American Literature (1880–2000)

Much of the recent work reviewed in section two studies the representation of gender in textual data. The currently dominant approach to this is to first use gendered anchor words (e.g., "man" and "woman," "boy" and "girl," "male" and "female") in order to estimate gender as a cultural dimension in a word embedding space (Kozlowski et al. 2019; Boutyline et al. 2020; Jones et al. 2020; Nelson 2021; for a notable exception to this approach, see Underwood 2019). In a second step, researchers then examine the correlation or cosine angle between the gender dimension and various other cultural dimensions, which is interpreted as semantic similarity or association. But what is the meaning of this measure of semantic association? What exactly does it tell us, for instance, if we find femininity rather than masculinity to be associated with affluence in the Google Ngram Corpus (Kozlowski et al. 2019:922–923)? According to the authors, this pattern is due to women historically serving as vessels for men's consumption. This explanation is plausible. However, it also points to the fact that relations uncovered by word embeddings can themselves be semantically underdetermined and that interpreting them remains a challenge.

The relations derivable with a semantic grammar are less susceptible to this problem, as their level of abstraction is considerably lower. Rather than how gender semantically aligns with various cultural dimensions or concepts, a semantic grammar allows us to capture how entities of different genders engage in narrative action and what attributes and characteristics are explicitly ascribed to them. The relations of a semantic grammar not only stay closer to the original text, but also to the conceptual ontologies of most sociological theories which, with few exceptions, involve stuff like actors, actions, and attributes. Simply put, instead of abstract semantic relations, we can capture representations of social relationships.

To demonstrate this, I use the U.S. Novel Corpus (USNC), a collection of 9,088 American novels published between 1880 and 1990 (Chicago-Text-Lab 2021; So, Long and Zhu 2019). Novels were selected based on the number of library holdings listed at Worldcat.org and therefore involve both mass-market and highly canonical publications (see supplemental material for additional corpus statistics). To identify gendered entities, I use the name register of the U.S. Social Security Administration which lists the frequency of all first names that were given to newborn children within the U.S. for any year between 1880 and 2000 (Wickham 2021). I use these to construct lists of given names that are indicative of either male

of female characters (see supplemental material for details). Furthermore, I consider all capitalized words preceded by "Mr.," "Mrs.," "Miss," and "Madame" as gendered entities, as well as the pronouns "he," "him," "his," "she," and "her." Next, the spaCy language pipeline is used for word tokenization, sentence splitting, and lemmatization as well as to annotate the text with part-of-speech tags and dependency trees. Finally, all sentences are processed with the *semgram* R package to extract motifs around tokens representing male and female entities.

For this analysis, I focus only on action-patient motifs, of which there are 7.5 million for male and 4.6 million for female core entities. In a next step, I take the subset of all action-patient motifs that have either a female or a male entity in the patient position (1.2 million). Distinguishing these by core entity gender and patient gender, I generate four sets of motifs: female-male actions, male-female actions, female-female actions, and male-male actions. I refer to male and female "characters" below although technically, there are some instances in which the entities referenced with gendered pronouns or names may not technically be human characters (for instance, animals, ships, or venues). Figure 5 provides a schematic representation of the workflow to this point.

Interestingly, most of the interaction in U.S. novels appears to occur across gender lines (see Table 3). Of all female action-patient motifs with gender-identified patient, 73.7% are directed at male characters. Of all male action-patient motifs with gender-identified patient, 69.3% are directed at female characters. In other words, an action by a gender-identified character in a novel is almost three times more likely to be directed at a character of the opposite gender than at a same-gender character. This is a considerable level of gender heterophily.

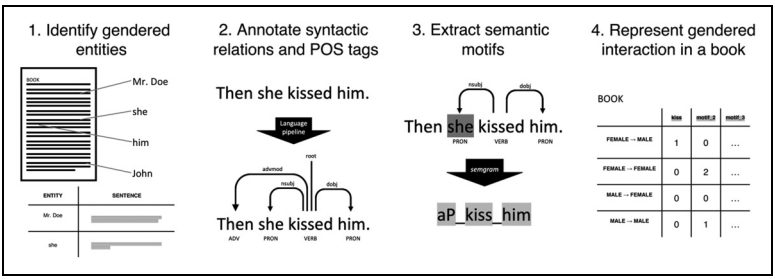

Figure 5. Workflow of analysis on gendered interaction in novels.

Table 3. Relative Frequencies of Same- and Inter-Gender Actions in U.S. Novels.

	Male patient	Female patient	Total
Male actions	17.3%	39.1%	56.4%
Female actions	32.2%	11.5%	43.7%
Total	49.5%	50.6%	

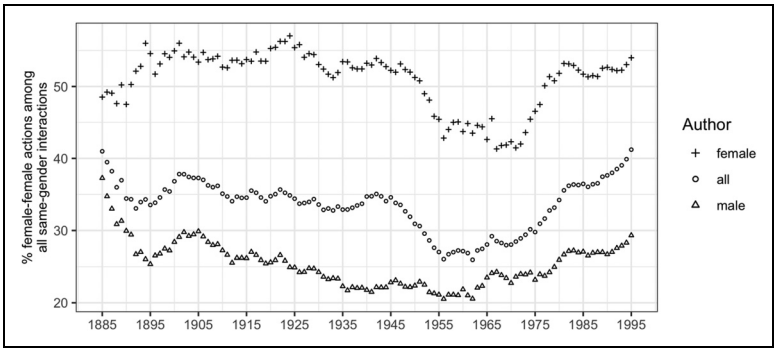

Figure 6. Share of female-female actions among same-gender interactions.
Note: Points show the percentage of female-female actions among all same-gender interactions averaged over books written within a 10-year time period. A value at 1950, for instance, represents all publications between 1946 and 1955.

A popular benchmark for evaluating the representation of women in works of fiction is the Bechdel-Test, which a movie or book passes if it contains two female characters who speak to each other about a topic other than a man. A surprisingly high number of works of fiction fail this basic criterion, including roughly half of the movies that have been named Best Picture at the Oscars (BBC 2018). While the grammar proposed here does not allow us to assess whether a book meets Bechdel's criterion, it does allow us to assess the relative frequencies of same-gender interaction. In the whole corpus female-female actions make up only 40.0% of all same-gender actions. Figure 6 shows the share of female-female actions among all same-gender interactions averaged over books and broken down by time and author gender. There is a relatively steady, accelerating decline from 1880 onwards with interactions among male characters becoming ever more prevalent compared to female-female interactions. Initially, this trend is driven by male, rather than female, authors. By the 1950s and 60s, the dominance of male-

male interactions reaches its peak and can even be found in novels written by women. From then on, there is a sharp reversal in this trend in both male- and female-authored books.

Nonetheless, a stable pattern throughout the whole period is that in books written by women female-female and male-male interactions are about equally prevalent. Men, on the other hand, are considerably more likely to write about male-male interactions. Throughout the 1950s and 60s, one was around 4 to 5 times as likely to find a male-male interaction in a novel written by a man than to find a female-female interaction.

Beyond the mere frequency, the grammar also allows us to investigate the semantics of gendered relations. First, I extract all actions for each of the four possible gender combinations: male-female, male-male, female-male, and female-female.[5] To find the actions most characteristic of a combination, I then compute the relative probability of each action within a combination. I use these to derive the probabilities of a combination conditional on each action, net of the combination's relative frequencies. Panel A of Figure 7 shows the actions most indicative of female-male and male-female combinations. To get an intuitive sense of these, consider the action "resist," which is associated with a probability of .62 for the female-male combination. This implies that if we were to observe an instance of resisting in a novel with equal motif counts of all combinations, there is a 62% chance of the sender being a female character and the recipient being a male character, rather than any of the other three possible combinations.

The actions most indicative of a male character acting towards a female one are related to sex ("penetrate," "bed," "enter"), violence ("rape," "crush"), and courtship ("court," "woo," "caress"). The motifs most indicative of a female character acting towards a male one, in turn, overwhelmingly have to do with resistance and rejection ("resist," "refuse," "repulse," "elude," "evade," "rebuff," "divorce," "defy"). Overall, the role relationship between male and female characters seems to be defined by the contrast between pursuit and resistance. Of course, we are only looking at the actions that are most characteristic for the respective gender combination, notwithstanding that there are prevalent modes of interaction between male and female characters that have nothing to do with this contrast. Panel B of Figure 7 shows the likelihood of different actions to be directed from a female character towards a male one, rather than vice versa. For instance, if there was an instance of "loving" between two characters of a different gender, it was about equally likely that a woman loved a man (.52) as it was that a man loved a woman. However, female characters are much more likely to "forgive" (.64) male ones. The same applies to "begging" (.65), "hating"

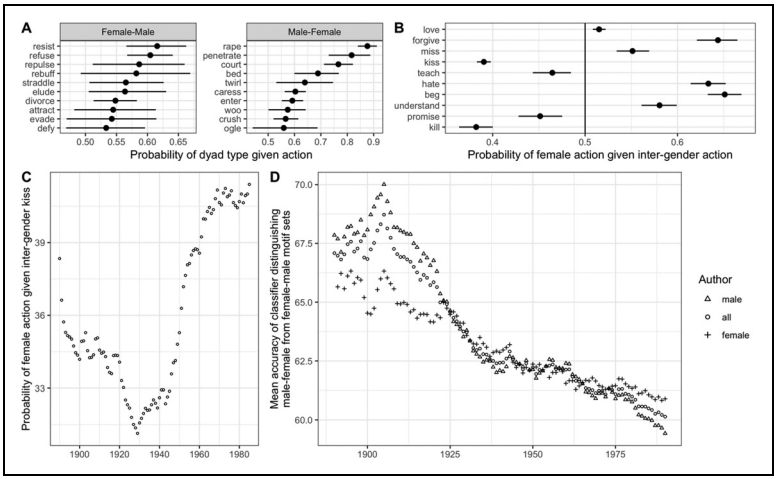

Figure 7. Content of cross-gender interaction in U.S. novels.
Note: Panel A shows the actions most characteristic of female-male and male-female combinations. Panel B shows the probability of a female-male combination (as opposed to a male-female one) for a selected set of actions. Panel C shows this statistic over time for the action of kissing. Points represent 25-year time windows. Panel D shows the accuracy of a naïve Bayes classifier distinguishing male-female from female-male motif sets as discussed in more detail in the text.

(.63), "understanding" (.58), and "missing" (.55). On the other hand, female characters are more likely to be "taught" (.46), "promised" (.45), or "killed" (.38) by male characters. There also appears to be a rather sizable kissing imbalance: 61% of all inter-gender kisses were received by female characters. Panel C investigates this imbalance over time and reveals a trend towards kissing equality during the 20th century, though this trend isn't linear.

Did gender relations generally become more symmetric? Asymmetry can be conceptualized as the degree to which actions are indicative of a particular gender combination. To operationalize this, I split the data into 20-year moving windows and consider only novels which contain at least ten female-male and ten male-female action-patient motifs, which applies to 75.1% of the corpus. For each time window, I sample 196 novels (95% of the number of novels contained in the window with the fewest novels). For every novel, I sample ten male-female and five female-male action motifs. I then use the 392 motif sets to train a naïve Bayes classifier for distinguishing male-female from female-male motif sets. This classifier is used to predict the gender

combination of all held-out motif sets. I repeat this procedure 1000 times for every time window in order to get robust accuracy estimates – even for the time windows with a comparatively small number of novels. Panel D in Figure 7 shows the average classification accuracy over those runs for each window. The classification accuracy can be interpreted as a diachronic measure of the asymmetry in gender relations. Note that the range of accuracy is a consequence of the artificial limit of only 10 motifs per motif set. The objective here is not to maximize prediction performance but to ensure comparability across time so that the scale of the reported accuracy is in some sense arbitrary.

Overall, there is an unambiguous trend towards symmetry. That is, the gender combination of action sets becomes less predictable over time, dropping from 68.8 around the turn of the 20th century to 60.0% in the latest time window between 1980 and 2000. Interestingly, the gender of the author appears to make little difference with regard to asymmetry. With a surprising level of concurrence, both male and female authors follow the general historical trend of degendering social interactions in their novels. One limitation of the analysis is that we collapse the motifs of a particular gender combination contained within a novel instead of distinguishing ties between individual characters, which might blur some of the gendered dynamics within novels. While addressing this lies beyond the scope of this paper, it might very well be worth pursuing in future research. Motif-based semantic grammar provides a new opportunity for studying representations of gendered interaction.

5.2 Who is "we"? Constructing Collective Identity in U.S. Presidential Campaigns (1952–2020)

When politicians speak at a campaign rally, they face a challenge: convincing the audience that they speak on their behalf. In the social movements context, this is known as "problem of identity correspondence" (Snow and McAdam 2000). Just like activists, politicians must attempt to construct a shared identity into which speaker and audience merge. The semantic reference points that lend themselves to this kind of identity work are highly varied and can include shared interests, values, enemies, history, destiny, class membership, nationality, ethnicity, or regional identity.

Here, I focus on U.S. presidential candidates' attempts to align their identities with those of the audience at campaign rallies. I do this by focusing on what is arguably the most important lexical anchor in this kind of identity

work: the first-person plural pronoun "we," together with its objective and possessive forms "us" and "our." Political sociologists have long noticed that pronouns play a key role in constructing and relating social identities (von Wiese 1965; Elias 1978), but don't seem to have explicitly attended to uses of "we" in political talk (though see Wagner-Pacifici 2010:1360-1361). Linguists and conversation analysts, however, have noticed its highly political character, for uses of "we" constitute demarcations of social boundaries. They are acts of simultaneous inclusion and exclusion (Tyrkko 2016; Bramley 2001; Pennycook 1994; Wodak et al. 2009:45). On the inclusive side, the "we" creates an association that usually involves speaker and audience but may well extend beyond that. On the exclusive side, uses of "we" in the political realm usually imply an "other," and can be used to stress "us"-"them" dichotomies.

To study the notions of "we" invoked by U.S. presidential candidates, I use a text corpus of campaign speeches and public statements created by Bonikowski and colleagues (2021). This data combines different sources to cover the elections from 1952 to 2020 and comprises 2,956 speeches delivered by 34 candidates. The data include speeches given by the two major parties' nominees between September 1 and Election Day, as well as their nomination acceptance speeches.

I extract from these texts all motifs around "we," "our," and "us" and represent them in lemmatized form. In total, there were 331,422 we-related motifs. The most frequent motif classes are actions, possessions, and patients (see Table 4). Agent and treatment motifs were considerably less likely, indicating that presidential candidates tend to construct an agentic "we" rather than a passive one. Note, furthermore, that the number of action-patient and agent-treatment motifs are equal to those of patient and agent motifs respectively, as any instance of the former implies an instance of the latter (see aP.2 in Table 2).

Table 4 lists the 5 most frequent motifs for each motif class. The top characterizations mostly ascribe collective strength and determination (be_strong, be_able, be_proud, be_sure). The most frequent possessions relate to "our" country and its people, as well as to the economy. The top action motifs are all semantically ambiguous, whereas the top patient motifs also refer to the country and its people (P_people, P_America, P_country), as well as to jobs (P_job). Action-patient motifs tend to be projective and related to political goals, such as creating jobs (aP_create_job), making progress (aP_make_progress), or simply winning the election (aP_win_election). The most frequent agent, treatment, and agent-treatment motifs are partially overlapping, indicating that there is not a lot of variation in the entities and

Table 4. Motifs in U.S. Presidential Campaigns.

Motif class	Total motifs	Top motifs	Clinton 2016	Trump 2016
Characterization	10277	be_strong, be_able, be_nation, be_proud, be_sure	be_well, be_good	be_rich
Possession	67868	H_country, H_people, H_nation, H_child, H_economy	H_election, H_work, H_Lady, H_difference, H_planet, H_commander, H_college	H_deal, H_deficit, H_politician, H_movement, H_love, H_shore
Action	141148	a_have, a_going-to, a_be, a_do, a_get	a_prove, a_listen, a_fear, a_trust, a_treat, a_accept	a_repeal, a_renegotiate, a_drain
Patient	49330	P_job, P_people, P_America, P_country, P_program	P_president, P_future, P_investment, P_million	P_deal, P_House, P_state, P_Obamacare, P_immigration, P_swamp, P_theft, P_crisis
Action-patient	49330	aP_win_election, aP_create_job, aP_make_progress, aP_build_bridge, aP_do_thing	aP_make_investment, aP_defend_right, aP_need_president, aP_grow_economy, aP_want_kind	aP_win_House, aP_win_state, aP_rebuild_military, aP_get_deal, aP_live_theft, aP_lower_tax, aP_replace_Obamacare
Agent	2294	A_people, A_God, A_attack, A_history, A_country		
Treatment	8881	t_give, t_tell, t_join, t_bring, t_help	t_call	
Agent-treatment	2294	At_God_give, At_attack_cost, At_Dante_tell, At_Bible_tell, At_history_teach		

actions towards "us." God giving us (At_God_give), the Bible telling us (At_Bible_tell), history teaching us (At_history_teach), and attacks costing us (At_attack_cost) are popular phrases among candidates.

To illustrate how candidates rely on different rhetorics in their construction of a collective identity, Table 4 also lists the 25 most characteristic[6] "we"-related motifs for the two 2016 campaigns. We can see, for instance, that Hillary Clinton's "we" is well (be_well) and good (be_good), whereas Donald Trump's is associated with being rich (be_rich). While Clinton seems to deliberately affiliate herself with the current administration when she speaks of "our" commander in chief (H_commander) or First Lady (H_Lady), Trump speaks of "our" politicians (H_politician) and draining (a_drain) the swamp (P_swamp) – thereby constructing an "us" that is set against elites. Fittingly, we find that in Trump's use, the referent of "we" frequently seems to be his own campaign or the Republican campaign (H_movement, P_House, aP_win_House, aP_win_state).

These examples then lead to a more general question: what are the dominant rhetorics that presidential candidates use in constructing a shared we-identity? To tackle this question, I build a series of motif-dictionaries. Word-dictionaries are subject to considerable skepticism in the social sciences (e.g., Grimmer and Stewart 2013:274–275). The main reason for this is that many of the qualities one might want to measure in a text aren't reducible to the mere presence or absence of particular words. Dictionaries built with motifs instead of words have the potential to overcome this pitfall because they contain clausal information that is constitutive of many textual qualities. Consider, for instance, that we observe the word "government" in proximity to the term "we." This in itself tells us relatively little about the framing of "we" as the semantic relation between the two concepts remains unspecified. Now consider, instead, that we know the word "government" corresponds to an instance of the motif "A_government." This tells us that the "government" is discussed as an agent treating "us" – in other words, the speaker constructs a collective identity that is defined *against* the government. In composite motifs such as "At_government_tell" or "At_government_treat" this framing becomes even more explicit. Granted, even motifs' semantics are subject to modification by the context and there are textual complexities that lie beyond their reach. For instance, the sentences "The government suppresses us" and "My opponent says, the government suppresses us" would both contain the motif "At_government_suppress," but only in the former is it a claim, whereas in the second, it is a second-order observation. Such shortcomings shouldn't obscure, however, that motifs generally allow us to enrich dictionaries with clausal information. They serve as

concentrates of claims contained within a text, which makes motif-based dictionaries potent measures of textual qualities.

To identify the dominant rhetorics around "we," I manually inspect the 75 most frequent motifs of each motif class – a total of 600 motifs (see documentation in Appendix D). To avoid bias against campaigns with fewer speeches, I select these based on their average probabilities across campaigns. While many of the most frequent motifs are semantically ambiguous (see Table 4), others appear to fall into particular registers used around "we." I identify five major rhetorics around "we" and assign individual motifs to them – leading to five dictionaries with an average number of 24 motifs. Table 5 shows the motifs associated with each rhetoric. Assignments were informed by the authors' familiarity with the speeches through previous projects, as well as surveying the instances in which specific motifs occurred. The prevalence of these rhetorics was then measured as the percentage of all "we"-related motifs belonging to the respective dictionary. These are shown in Figure 8.

One rhetoric frequently employed by presidential candidates posits "we" in opposition to the government or current administration. Often, this is done by describing how "we" are treated by the government as an *agent* (A_President, A_administration, A_government, A_president). The most common phrase is that "we" are being "told" things by the government, as seen in the respective *agent-treatment* motifs (At_President_tell, At_government_tell, At_administration_tell). Besides, a number of *action-patient* motifs stress requirements for an administration (aP_need_government, aP_need_president, aP_want_president), which is usually a way of emphasizing the present administration's incompetence. Another popular way to phrase this is that we cannot afford more years of the current administration (aP_afford_year), or that we are tired (be_tired) of current circumstances. "We"-versus-administration rhetoric is employed by both parties alike. Use depends primarily on who is a challenger to an incumbent president or party (correlation of .60, $p < .001$). The three campaigns relying most heavily on such rhetoric were those of Mondale in 1984, Dukakis in 1988, and Obama in 2008 – all of whom ran against Republicans after years of Republican incumbency.

A "we" can also be constructed by emphasizing a common security-political identity and evoking an external threat. "We" are attacked (t_attack), hurt (t_hurt), or hit (t_hit) by "our" enemies (H_enemy, A_enemy) or adversaries (A_adversary) and handle conflicts (P_peace, P_war) by means of "our" military forces (H_troop, H_military, H_force). Using "we" in discussing such matters, rather than, say, "the United

States," merges speaker and audience into a national community of fate. Generally, Republican candidates are more likely to employ such motifs (correlation of .39, $p < .05$). However, the use of military/threat-rhetoric also depends on the historic circumstances of a campaign. By far, the highest prevalence of this rhetoric occurred in 2004, the first presidential election after the September 11 attacks and the invasion of Iraq.

The most prevalent rhetoric around "we" concerns the United States as an economy and fiscal unit. Candidates talk about "our" economy, jobs, and businesses (H_economy, H_job, H_business), as well as how "we" can affect it (e.g., aP_create_job, aP_cut_taxes, aP_grow_economy). Use of this rhetoric is characterized by an interesting historical trend: talk about "our" economy wasn't particularly prevalent up until the beginning of the 1980s when such rhetoric started to become ever more popular among candidates. It reached its peak in the 2008 campaigns during the financial crisis, making up 8 to 9% of all "we"-related motifs. Since then and with the recovery of the U.S. economy, its prevalence has leveled off again.

It is also common for candidates to address "us" as family members. The claim here isn't that speaker and audience share one family. Rather, speaking of "our" children and family members (H_child, H_family, H_kid, H_grandchild) is a way of stressing people's shared roles as parents and members of families broadly. With only two recent exceptions (Bush 2004 and Romney 2012), this rhetoric tends to be more popular among Democratic candidates. It was by far the most prevalent in Gore's 2000 campaign. Further investigation reveals that Gore made use of this rhetoric because through identifying "we" with the family, he was able to talk about the seemingly distant consequences of climate change.

Finally, another use of "we" is to equate it with the campaign itself. Candidates speak about "our" campaign, party, or opponent (H_opponent, H_party, H_campaign), discuss how the other side treats "us" (A_Republicans, A_Democrats, A_opponent, At_opponent_tell), or about how "we" fare in the election (a_win, aP_win_election, aP_take_office). Identifying collective identity with one's campaign appears to be somewhat idiosyncratic. It is neither significantly associated with a party, nor is there a challenger-incumbent dynamic or a historic trend. In recent years, this form of "we"-rhetoric has been most prevalent in Trump's 2016 and 2020 campaigns, as well as in Obamas 2008 and 2012 campaigns. Historical precedents are Nixon's 1960 campaign against Kennedy and Mondale's 1984 campaign against Nixon. That said, further investigation of the data points to an interesting dynamic: the degree to which candidates identify "we" with their own campaign appears to shift over the course of the campaign.

Table 5. "we"-Rhetorics and Corresponding Motif Dictionaries.

	Opposition to administration	Campaign	Threatened/military	Family	Economy/Fiscal unit
Agent	A_President, A_administration, A_government, A_president	A_Republicans, A_opponent, A_Democrats, A_party, A_poll	A_attack, A_enemy, A_force, A_crisis, A_Communists, A_adversary	A_family, A_senior	
Treatment		t_elect	t_divide, t_serve, t_attack, t_hurt, t_destroy, t_hit, t_overtake, t_fight		t_cost
Action		a_win			a_spend, a_afford, a_invest
Patient	P_president, P_government	P_election	P_peace, P_war, P_force	P_family, P_child, P_school	P_job, P_economy, P_taxes, P_money, P_business, P_budget, P_worker, P_cost, P_deficit, P_market, P_dollar
Characterization		be_party	be_safe, be_secure, be_divided	be_family	be_prosperous
Possession		H_opponent, H_party, H_campaign, H_debate	H_troop, H_security, H_military, H_force, H_defense, H_peace, H_enemy	H_child, H_family, H_school, H_kid, H_grandchild	H_economy, H_job, H_worker, H_prosperity, H_deficit, H_market, H_business
Agent-treatment	At_President_tell,	At_opponent_tell,	At_enemy_fear,	At_senior_raise	

(continued)

Table 5. Continued

	Opposition to administration	Campaign	Threatened/military	Family	Economy/Fiscal unit
	At_government_tell, At_administration_give, At_president_put, At_administration_tell, At_president_tell	At_Republicans_join, At_Republicans_give, At_poll_show, At_opponent_send, At_opponent_give	At_attack_cost, At_force_divide, At_crisis_teach		
Action-patient	aP_need_president, aP_afford_year, aP_need_government, aP_want_president, aP_need_President	aP_win_election, aP_take_office, aP_win_victory, aP_win_state	aP_keep_peace, aP_end_war, aP_want_peace, aP_fight_war		aP_create_job, aP_cut_taxes, aP_balance_budget, aP_spend_money, aP_open_market, aP_add_job, aP_lose_job, aP_bring_job, aP_grow_economy, aP_build_economy, aP_reduce_deficit

As Election Day approaches, campaign identity seems to play an ever-increasing role in candidates' rhetoric. The bottom right panel in Figure 8 shows the correlation between the prevalence of the "campaign"-rhetoric and the number of days until the election. There is a negative association for 8 of the 10 campaigns since 2000, though not all these correlations are statistically significant. This finding aligns with research on negative campaigning showing that candidates increase their emphasis on campaign boundaries as elections approach (Damore 2002; De Nooy and Kleinnijenhuis 2013). At the outset of a campaign, candidates tend to emphasize issue content in order to inform the electorate who they are and what is important to them. As elections become imminent and candicates have successfully built their profiles, "we" more often becomes a marker for a candidates campaign in explicit juxtaposition to the opponent's.

Overall, this analysis illustrates the potential that lies in basing dictionaries on motifs, rather than the mere occurrence of words. By integrating clausal information into our measurement tool, relatively complex textual qualities become measurable. Here we were able to measure different rhetorics serving the construction of collective identity – a phenomenon likely to evade many other measurement strategies. The analyses also illustrate another central advantage of the framework important to sociologists: whereas many contemporary machine learning methods tend to operate as de facto black boxes, a representation via semantic motifs stays close to the original text. It provides transparency that opens measurement up for critique and scrutiny.

6. Validation and Discussion of Limitations

To assess the quality of motif extraction, I draw two samples from the data used in the analyses: Sample 1 consists of 250 sentences from the USNC that contain gendered entities, stratified by decade; Sample 2 consists of 250 sentences from the Campaign Speeches Corpus which contain "we," "us," or "our," stratified by campaign. For each sentence, I manually annotate and list all motifs around the core entities. These lists serve as the gold standard and are compared with the motifs that were computationally extracted. To make this process as transparent as possible, the two annotated datasets are provided in the supplemental material, together with comments and elaborations on individual annotations.

Tables 6 and 7 show the results for each motif class. The recall value represents the portion of true motifs that were computationally extracted. For instance, of the 778 motifs that were annotated in the campaign speech

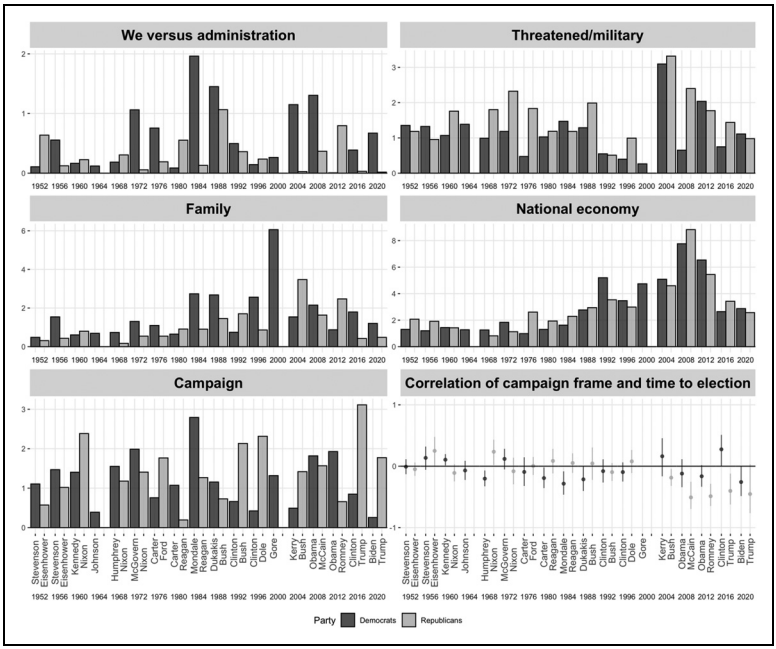

Figure 8. Prevalence of "we"-rhetorics in presidential campaigns.
Note: The bar plot panels show the prevalence of different rhetorics as percentages of all "we"-related motifs that belong to the respective dictionary. For instance, 2.57% of all "we"-related motifs employed in the Trump 2020 campaign belong to the national economy frame. The bottom right panel shows the correlation between the prevalence of the "campaign"-rhetoric and the logarithmized number of days until the election with 90% confidence intervals. The 2000 Gore campaign was removed from this panel because there are only 4 speeches in the data.

sample, 652 were successfully extracted, leading to an overall recall of .84. For the USNC sample, recall reaches .80. Our ability to extract motifs differs with motif class. Whereas almost all action (.90 and .82), treatment (.95 and .82), and possession (.96 and .99) motifs could be successfully extracted, the performance with regard to patients (.76 and .72), agents (.47 and .71), and the two composite motif classes is considerably lower. This is likely due to the fact that these motifs are usually more distant to the entities of interest in the dependency graphs. In fact, the identification of agents and patient depends on the correct identification of actions and treatments, making the former a more complex task. Finally, characterization motifs show a

Table 6. Precision and Recall for Campaign Speeches Validation Sample.

	Actions	Patients	Action-patients	Treatments	Agents	Agent-treatments	Characterizations	Possessions	Overall
Recall	0.90	0.76	0.76	0.95	0.47	0.47	0.92	0.96	0.84
Precision	0.94	0.87	0.87	1	0.90	0.90	0.77	0.92	0.90
N true	262	147	147	20	19	19	25	139	778
N extracted	252	128	128	19	10	10	30	145	722

Table 7. Precision and Recall for USNC Sample.

	Actions	Patients	Action-patients	Treatments	Agents	Agent-treatments	Characterizations	Possessions	Overall
Recall	0.82	0.72	0.72	0.82	0.71	0.71	0.65	0.99	0.80
Precision	0.96	0.89	0.89	0.80	0.88	0.88	0.97	0.96	0.93
N true	293	130	130	40	31	31	51	136	842
N extracted	250	104	104	41	25	25	34	140	723

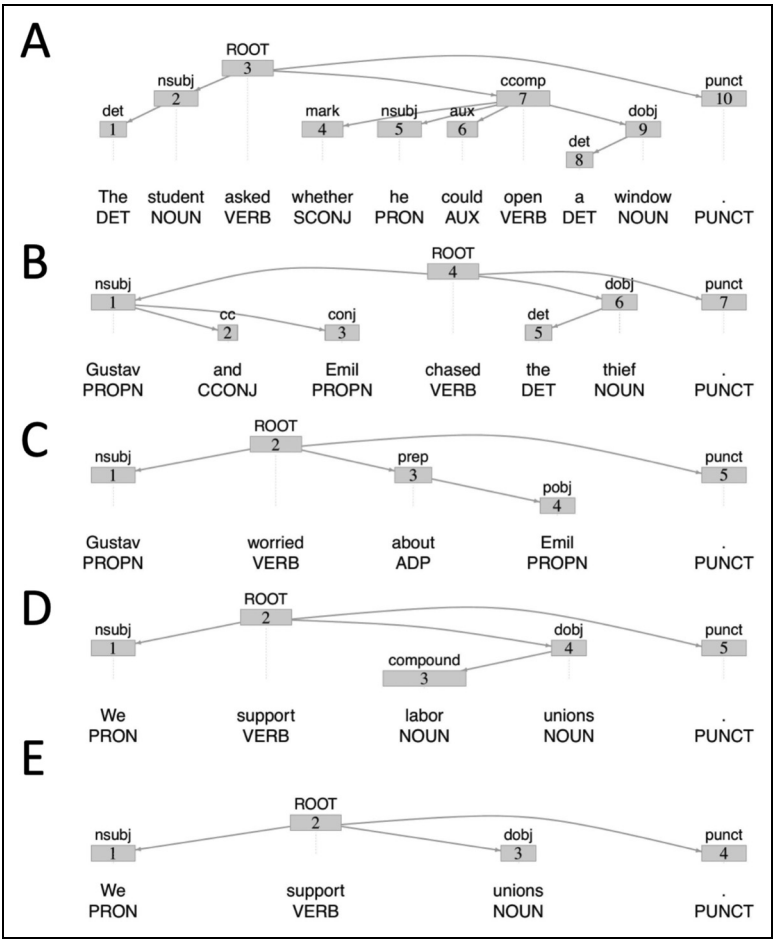

Figure 9. Exemplary sentences illustrating complexities in motif extraction via syntactic rules.

somewhat distinguished pattern with a recall of .65 in the USNC sample and .92 in the Campaign Speeches sample. Overall, recall values indicate that there is room for improvement in motif extraction and that a non-negligible amount of relevant information contained within texts gets lost in this process.

However, a perhaps more important measure regarding the validity of the analyses is precision – that is, the portion of extracted motifs that are true motifs. Precision is .90 for the Campaign Speeches corpus and .93 for the USNC. We again find similar differences among motif classes, which are, however, less pronounced than in recall. In nearly all motif classes, we find that about 9 out of 10 extracted motifs are correct. The one exception to this is the characterization motifs in the Campaign Speech sample, indicating that the quality of extraction can differ depending on the language style.[7]

These numbers point to a general pattern: our ability to extract motifs leaves some room for improvement, especially with regard to the dependent motif classes *agent* and *patient*. However, the motif that were extracted are overwhelmingly correct. To further investigate potential biases in the quality of motif extraction, I ran a series of multilevel regression analyses on sentence and document level covariates which are provided in the supplemental material. Given that the spaCy language pipeline used here was largely trained on textual data from contemporary sources, it seemed especially important to assess whether extraction performance changes over time. I find some evidence suggesting that a higher number of extractions in a sentence increases the likelihood of erroneous extractions (precision). However, neither the year of the presidential campaign nor the year of a novel's publication are significant predictors of extraction performance. The same applies to sentence length and the number of true motifs (that is, human annotated motifs) within a sentence. Together, these results suggest that the approach proposed here is capable of extracting valid motifs across a variety of applications. Besides, whether or not heterogeneity in extraction performance is a cause for worry depends on the goals of the subsequent analysis. While it seems likely that extraction performance will differ between vastly different tasks and texts, none of the validation efforts suggest particular vulnerability to such variations.

That said, it is important to acknowledge that there is relevant semantic information within textual data which simply lies beyond what the approach proposed here can extract in its current form. This is due to a series of limitations. The first, and probably most important of these, relates to the phenomenon of coreference; often, multiple words or phrases stand in for the same entity. To illustrate this, consider sentence A in Figure 9. A first reading of the sentence might suggest that "student" and "he" refer to the same person. Assuming that we were interested in extracting motifs around gendered entities, we would ideally want to extract "a_ask" as an action motif, given that the "student" and "he" are the same person. However, the acting entity associated with "asked" isn't "he" but "student" – which is itself not

a gendered entity, and so, our approach would have failed to extract "a_ask." Coreference is frequent in natural language and often spans sentence boundaries. There is no question that in many applications, it causes us to miss relevant information. Sentence A, however, also lends itself to illustrate the intricacy of coreference: isn't it equally plausible that the sentence is about a student of unknown gender asking, for instance, a male teacher whether "he" could open a window? There is, quite simply, no way to know for sure, especially without drawing on contextual information. Coreference resolution – the task of identifying all words and phrases referring to the same entity – is difficult because it involves complex interpretations. It is an active research area in natural language processing. Future versions of the approach proposed here may integrate a coreference resolution component. However, this will inevitably come at the cost of reduced precision, which may not be desirable for many social science applications.

Second, the grammar proposed here relates core entities to the patients it directly acts upon and to the agents that directly act upon it. It does this if, and only if, patients and agents are linked to the core entity through a transitive verb. There are, however, many ways in which texts inform about the relationship between entities that do not follow this form. In sentence B of Figure 9, for instance, Gustav and Emil engage in a common action – something that the approach proposed here would currently not capture. Similarly, Sentence C informs us that Gustav worried about Emil. However, instances in which the relationship between entities are mediated by a preposition ("about"), are not currently considered for the grammar. Doing so would require either expanding the notion of motifs to include multi-word phrases (e.g., a_worry_about), or defining new motif classes. Increasing the complexity of the grammar in this way may be the focus of future work. There will, however, always be information about relationships between entities that an approach based on syntax cannot capture.

This leads us to a third limitation: the basic unit of the grammar is the token. In some cases, however, this could lead to the discarding of relevant information. For example, if we were to extract motifs associated with "we" in sentence D, we would extract "P_union" and "aP_support_union," even though it may be relevant what kinds of unions are supported. Therefore, it might seem attractive to use phrases such as "labor unions" as motifs, rather than single tokens. The pragmatic reason for not using multi-word phrases as motifs is that this would likely introduce sparsity into the extracted data that is unhelpful for most subsequent analyses. This is because phrases corresponding to the same referent often take only slightly different forms, as illustrated by sentence E. Here, support is directed at

"unions" alone. Using phrases, rather than tokens, would create semantically congruent but lexically distinct motifs and thereby significantly increase the size of the motif vocabulary. There is clearly a tradeoff here between the reduction of sparsity and the retention of potentially relevant information. By using tokens as basic units of motifs, we pursue the former at the cost of sacrificing some of the latter.

Finally, the approach presented here does not capture negation. Apart from the fact that integrating negation would further complicate the grammar, it also turns out to be hard to implement. Difficulty arises here primarily from the fact that syntactic patterns don't always seem to provide sufficient information for deciding whether a negation applies transitively to a chain of verbs or adjectives. While a disregard for negation is the norm among the text analysis approaches currently used in sociology (see Section 2), it may be worthwhile to implement a naive approximation into the grammar in future work. In fact, none of the various limitations discussed here should be seen as irresolvable. Instead, they are the construction sites where future efforts may take off and refine the basic framework introduced here.

7. Conclusion

Sociologists have formally analyzed semantic relations in textual data since the late 1980s. While this research program has recently gained attention and found its way into the discipline's mainstream, it has also become dominated by studies that focus either on co-occurrences or on semantic distances between concepts. The framework outlined here provides an alternative to these approaches that has at least three advantages.

First, the framework makes clausal information accessible to measurement. Many of the things that sociologists care for in a text cannot be reduced to the presence or absence of particular words. By integrating clausal information into our measurement tools, we increase our capacity to capture relatively complex textual qualities. Of course, as has been acknowledged, there is a limit to the argumentative complexity that can be represented by motifs. That is why this advancement should realistically be described as a gradual one: with a semantic grammar, we can move beyond what a text *is about* and towards what it *claims*; beyond the measurement of *themes and topics* and towards the measurement of *arguments*.

Second, the relative proximity of its representational form to the original text promotes intersubjectively valid interpretations. No formal approach frees us from having to make sense of the patterns it uncovers. But the currently dominant representational forms tend to abstract away the events and narrative

action that takes place within texts, thereby leaving a lot of ground to be covered by means of interpretation. What does it mean if we find that concepts co-occur or have similar word embeddings? Too often, there seem to be multiple plausible answers, and the criteria for selecting among them aren't obvious. If, on the other hand, we find that male novel characters kiss female novel characters more often than vice versa, this finding can stand on its own, requiring comparatively little interpretation. To be clear, this is not to say that such a result isn't premised on ex-ante decisions, nor that it needs *no* interpretation, critical assessment, and follow-up questions. I do maintain, however, that the gap between formal pattern and articulable, intersubjectively valid sociological insight is considerably smaller.

This then leads to what is probably the most important advantage of the proposed approach: its proximity to the ontology of sociological theory. In this framework, the concepts we conventionally use in our descriptions of social structures – actors, interactions, relationships, roles, and attributes, to name a few – become measurable in text. To this end, section 5.1 analyzed *interactions* between gendered characters; section 5.2 analyzed the construction of "us" as an *actor* in political rhetoric; and elsewhere, it was recently shown that a similar approach can be used to measure *roles* in textual data (Stuhler 2021). Not least, this paper is therefore an attempt to provide tools and contours for what could become a new research program. This program moves beyond studying abstract textual associations and towards a formal analysis of *textual representations of social structures*.

Acknowledgments

I am grateful to Delia Baldassarri, Bart Bonikowski, Adam Braffman, Paul DiMaggio, Jan Fuhse, and Patrick Kaminski, as well as to the members of the Culture and Action Network workshop at the University of Chicago for their suggestions and comments on previous drafts. Special thanks go to Hoyt Long for granting me access to the U.S. Novel Corpus.

Author's Note

A repository providing replication materials and supplementary files can be accessed under the link printed below. Detailed descriptions of files and instructions for replicating the analyses can also be found in the repository. https://osf.io/rq6j7/?view_only=52bccf58ea004439853dff01f464ac50

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Oscar Stuhler https://orcid.org/0000-0001-7391-1743

Supplementary Material

Supplemental material for this article is available online.

Notes

1. While this might seem like an odd choice, "patient" is the common term used in semantic grammars to denote entities that are acted upon.
2. Note that spaCy offers multiple processing pipelines which provide different trade-offs between performance and computational cost. Here, I use the largest model for optimal performance. For specifics on the pipeline, see https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.1.0
3. While I make a sharp distinction here between *syntactic* and *semantic* grammars, it is worth acknowledging that the precise boundary between these two categories is not always obvious and that modern dependency grammars, while arguably focused on syntax, take into account some semantic criteria (see the discussion in de Marneffe and Nivre 2019:203). While the distinction may thus in fact be more gradual than categorical, it is nonetheless important to highlight the gap that exists between primarily *syntax-focused*, fine-grained dependency grammars and the kind of *semantic* grammar that is proposed in this paper.
4. The current version of *semgram* as well as a demo can be found at https://github.com/omstuhler/semgram
5. The motifs were analyzed in lemmatized form – meaning that lexical features were reduced to their base form. This smoothens the presentation of the results but also helps to make the data less sparse. However, one may well choose to study unlemmatized motifs given that these can contain information relevant for analytic goals other than mine such as, for instance, tense.
6. Motifs were selected according to a frequency-exclusivity score. The score is the mean of a motif's (m_i), rank by probability under a particular candidate (c_j), that

is $P(m_i|c_j)$ and its rank by exclusivity under a given candidate, where exclusivity is defined as $P(c_j|m_i)$ after normalizing for unequal amounts of data per candidate.

7. Further inspection of this particular result led to no clear explanation. There is, however, an unusually high number of extractions in the Campaign Speech sample compared to the gold standard annotation (20% more extractions than true motifs). Inspection of the false positive cases suggests that this is a consequence of the way in which candidates speak about what "we are:" often sentences start with "we" and a copula verb (e.g., "we are"), followed by multiple subordinate clauses. False positives tend to be located in these subordinate clauses.

References

Abell, Peter. 1987. *The Syntax of Social Life: The Theory and Method of Comparative Narratives*. Oxford: Oxford University Press.

Abell, Peter. 2004. "Narrative Explanation: An Alternative to Variable-Centered Explanation?" *Annual Review of Sociology* 30(1):287-310. doi:10.1146/annurev.soc.29.010202.100113

Arseniev-Koehler, Alina. 2021. "Theoretical Foundations and Limits of Word Embeddings: What Types of Meaning can They Capture?. https://arxiv.org/abs/2107.10413

Arseniev-Koehler, Alina and Jacob Foster. 2020. "Machine Learning as a Model for Cultural Learning: Teaching an Algorithm What it Means to be fat." *https://arxiv.org/abs/2003.12133*.

BBC. 2018. "100 Women: How Hollywood fails women on screen." https://www.bbc.com/news/world-43197774.

Bearman, Peter, Robert Faris, and James Moody. 1999. "Blocking the Future: New Solutions for Old Problems in Historical Social Science." *Social Science History* 23(4):501-33. doi:10.1017/S0145553200021854

Bearman, Peter and Katherine Stovel. 2000. "Becoming a Nazi: A Model for Narrative Networks." *Poetics* 27(2-3):60-90.

Benoit, Kenneth and Akitaka Matsuo. 2020. "spacyr: An R wrapper to the Python spaCy NLP library." https://cran.r-project.org/web/packages/spacyr/index.html.

Bonikowski, Bart, Yuchen Luo, and Oscar Stuhler. 2021. "Politics as Usual? Antecedents of Radical-Right Frames in U.S. Electoral Discourse." https://doi.org/10.31235/osf.io/uhvbp.

Boutyline, Andrei, Alina Arseniev-Koehler, and Devin J Cornell. 2020. "School, Studying, and Smarts: Gender Stereotypes and Education Across 80 Years of American Print Media, 1930-2009." https://doi.org/10.31235/osf.io/bukdg.

Bramley, Nicolette Ruth. 2001. "Pronouns of politics: the use of pronouns in the construction of 'self' and 'other' in political interviews." Series Pronouns of politics:

the use of pronouns in the construction of 'self' and 'other' in political interviews., Edition. Australian National University: Publisher.

Carley, Kathleen. 1988. "Formalizing the Social Expert's Knowledge." *Sociological Methods & Research* 17(2):165-232. doi:10.1177/0049124188017002003

Carley, Kathleen. 1993. "Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis." *Sociological Methodology* 23:75-126. doi:10.2307/271007

Carley, Kathleen. 1994. "Extracting Culture Through Textual Analysis." *Poetics* 22(4):291-312. doi:10.1016/0304-422X(94)90011-6

Carley, Kathleen and Michael Palmquist. 1992. "Extracting, Representing, and Analyzing Mental Models." *Social Forces* 70(3):601-36. doi:10.2307/2579746

Chatman, Seymour B. 1978. *Story and Discourse. Narrative Structure in Fiction and Film*. Ithaca: Cornell University Press.

Chicago-Text-Lab. 2021. "U.S. Novel Corpus." https://textual-optics-lab.uchicago.edu/us_novel_corpus.

Choi, Jinho D. and Martha Palmer. 2012. Guidelines for the Clear Style Constituent to Dependency Conversion. https://www.mathcs.emory.edu/~choi/doc/cu-2012-choi.pdf

Damore, David F. 2002. "Candidate Strategy and the Decision to Go Negative." *Political Research Quarterly* 55(3):669-85. doi:10.1177/106591290205500309

Danowski, James A. 1993. "Network Analysis of Message Content." *Progress in Communication Sciences* 12:197-221.

de Marneffe, Marie-Catherine and Christopher D. Manning. 2008. "Stanford typed dependencies manual." https://downloads.cs.stanford.edu/nlp/software/dependencies_manual.pdf.

de Marneffe, Marie-Catherine and Joakim Nivre. 2019. "Dependency Grammar." *Annual Review of Linguistics* 5(1):197-218. doi:10.1146/annurev-linguistics-011718-011842

De Nooy, Wouter and Jan Kleinnijenhuis. 2013. "Polarization in the Media During an Election Campaign: A Dynamic Network Model Predicting Support and Attack Among Political Actors." *Political Communication* 30(1):117-38. doi:10.1080/10584609.2012.737417

Deerwester, Scott, Susan T. Dumais, George W. Furnas, and Thomas K. Landauer. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41(6):391-407.

DiMaggio, Paul, 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. doi:10.1016/j.poetic.2013.08.004

Duquenne, Vincent, John W. Mohr, and Annick Le Pape. 1998. "Comparison of Dual Orderings in Time." *Social Science Information* 37(2):227-53. doi:10.1177/053901898037002001

Eisenstein, Jacob. 2019. *Introduction to Natural Language Processing*. Cambridge: MIT Press.

Elias, Norbert. 1978. *What is Sociology*. New York: Columbia University Press.

Fligstein, Neil, Jonah Stuart 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. doi:10.1177/0003122417728240

Franzosi, Roberto. 1989. "From Words to Numbers: A Generalized and Linguistics-Based Coding Procedure for Collecting Textual Data." *Sociological Methodology* 19(1):263-98. doi:10.2307/270955

Franzosi, Roberto. 1990. "Computer-Assisted Coding of Textual Data." *Sociological Methods and Research* 19(2):225-57. doi:10.1177/0049124190019002004

Franzosi, Roberto. 1994. "From Words to Numbers: A Set Theory Framework for the Collection, Organization, and Analysis of Narrative Data." *Sociological Methodology* 24:105-36. doi:10.2307/270980

Franzosi, Roberto. 1998a. "Narrative Analysis - Or why (and how) Sociologists Should be Interested in Narrative." *Annual Review of Sociology* 24(1):517-54. doi:10.1146/annurev.soc.24.1.517

Franzosi, Roberto. 1998b. "Narrative as Data: Linguistic and Statistical Tools for the Quantitative Study of Historical Events." *International Review of Social History* 43(6):81-104. doi:10.1017/S002085900011510X

Franzosi, Roberto. 2009. *Quantitative Narrative Analysis*. Thousand Oaks: Sage Publications.

Fuhse, Jan, Oscar Stuhler, Jan Riebling, and John Levi Martin. 2020. "Relating Social and Symbolic Relations in Quantitative Text Analysis. A Study of Parliamentary Discourse in the Weimar Republic." *Poetics* 78. doi:10.1016/j.poetic.2019.04.004

Goldenstein, Jan and Philipp Poschmann. 2019. "Analyzing Meaning in Big Data: Performing a Map Analysis Using Grammatical Parsing and Topic Modeling." *Sociological Methodology* 49(1):83-131. doi:10.1177/0081175019852762

Grimmer, Justin and Brandon M Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267-97. doi:10.1093/pan/mps028

Hoffman, Mark Anthony, Jean-Philippe Cointet, Philipp Brandt, Newton Key, and Peter Bearman. 2018. "The (Protestant) Bible, the (Printed) Sermon, and the Word(s): The Semantic Structure of the Conformist and Dissenting Bible, 1660–1780." *Poetics* 68:89-103. doi:10.1016/j.poetic.2017.11.002

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. "spaCy: Industrial-strength Natural Language Processing in Python." https://spacy.io.

Jones, Jason, Mohammad Amin, Jessica Kim, and Steven Skiena. 2020. "Stereotypical Gender Associations in Language Have Decreased Over Time." *Sociological Science* 7:1-35. doi:10.15195/v7.a1

Jurafsky, Dan and James H Martin. 2020. "Chapter 14. Dependency Parsing." *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/14.pdf.

Kang, Donghyun and James Evans. 2020. "Against Method: Exploding the Boundary Between Qualitative and Quantitative Studies of Science." *Quantitative Science Studies* 1(3):930-44. doi:10.1162/qss_a_00056

Karell, Daniel and Michael Freedman. 2019. "Rhetorics of Radicalism." *American Sociological Review* 84(4):726-53. doi:10.1177/0003122419859519

Koopmans, Ruud and Paul Statham. 1999. "Political Claims Analysis: Integrating Protest Event and Political Discourse Approaches." *Mobilization* 4(2):203-21. doi:10.17813/maiq.4.2.d759337060716756

Kozlowski, Austin C., Matt Taddy, and James A Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class Through Word Embeddings." *American Sociological Review* 84(5):905-49. doi:10.1177/0003122419877135

Lee, Monica and John Levi Martin. 2014. "Coding, Counting and Cultural Cartography." *American Journal of Cultural Sociology* 3(1):1-33.

Lee, Monica and John Levi Martin. 2018. "Doorway to the Dharma of Duality." *Poetics* 68:18-30. doi:10.1016/j.poetic.2018.01.001

Leydesdorff, Loet and Iina Hellsten. 2006. "Measuring the Meaning of Words in Contexts: An Automated Analysis of Controversies About 'Monarch Butterflies,' 'Frankenfoods,' and 'stem Cells'." *Scientometrics* 67(2):231-58. doi:10.1007/s11192-006-0096-y

Light, Ryan. 2014. "From Words to Networks and Back." *Social Currents* 1(2):111-29. doi:10.1177/2329496514524543

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Berthard, and David McClosky. 2014. "The Stanford CoreNLP Natural Language Processing Toolkit." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*: 55-60. doi:10.3115/v1/P14-5010

Martin, John Levi. 2000. "What Do Animals Do all Day?: The Division of Labor, Class Bodies, and Totemic Thinking in the Popular Imagination." *Poetics* 27-(2-3):195-231. doi:10.1016/S0304-422X(99)00025-X

McLean, Paul D. 1998. "A Frame Analysis of Favor Seeking in the Renaissance: Agency, Networks, and Political Culture." *American Journal of Sociology* 104(1):51-91. doi:10.1086/210002

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv:1301.3781*.

Mische, Ann and Philippa E Pattison. 2000. "Composing a Civic Arena: Publics, Projects, and Social Settings." *Poetics* 27(2-3):163-94. doi:10.1016/S0304-422X(99)00024-8

Mohr, John W. 1994. "Soldiers, Mothers, Tramps and Others: Discourse Roles in the 1907 New York City Charity Directory." *Poetics* 22(4):327-57. doi:10.1016/0304-422X(94)90013-2

Mohr, John W. and Petko Bogdanov. 2013. "Introduction—Topic Models: What They are and why They Matter." *Poetics* 41(6):545-69. doi:10.1016/j.poetic.2013.10.001

Mohr, John W. and Vincent Duquenne. 1997. "The Duality of Culture and Practice: Poverty Relief in New York City, 1888-1917." *Theory and Society* 26(2/3):305-56. doi:10.1023/A:1006896022092

Mohr, John W. and Helene K Lee. 2000. "From Affirmative Action to Outreach: Discourse Shifts at the University of California." *Poetics* 28(1):47-71. doi:10.1016/S0304-422X(00)00024-3

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. doi:10.1016/j.poetic.2013.08.003

Monroe, Burt L. 2019. "The Meanings of "Meaning" in Social Scientific Text Analysis." *Sociological Methodology* 49(1):132-9. doi:10.1177/0081175019865231

Nelson, Laura K. 2021. "Leveraging the Alignment Between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century U.S. South." *Poetics* 88. doi:10.1016/j.poetic.2021.101539

Padgett, John F., Katalin Prajda, Benjamin Rohr, and Jonathan Schoots. 2020. "Political Discussion and Debate in Narrative Time: The Florentine Consulte e Pratiche, 1376–1378." *Poetics* 78. doi:10.1016/j.poetic.2019.101377

Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. "GloVe: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*: 1532-43. doi:10.3115/v1/D14-1162

Pennycook, Alastair. 1994. "The Politics of Pronouns." *ElT Journal* 48(2):173-8. doi:10.1093/elt/48.2.173

Polletta, Francesca, Pang Ching Bobby Chen, Beth Gharrity Gardner, and Alice Motes. 2011. "The Sociology of Storytelling." *Annual Review of Sociology* 37(1):109-30. doi:10.1146/annurev-soc-081309-150106

Propp, Vladimir. [1928] 1968. *Morphology of the Folktale*. Austin: University of Texas Press.

Puetz, Kyle, Andrew P. Davis, and Alexander B Kinney. 2021. "Meaning Structures in the World Polity: A Semantic Network Analysis of Human Rights Terminology in the World's Peace Agreements." *Poetics*. 88. doi:10.1016/j.poetic.2021.101598

Roberts, Carl W. 1989. "Other Than Counting Words: A Linguistic Approach to Content Analysis." *Social Forces* 68(1):147-77. doi:10.2307/2579224

Roberts, Carl W. 1997. "A Generic Semantic Grammar for Quantitative Text Analysis: Applications to East and West Berlin Radio News Content from 1979." *Sociological Methodology* 27(1):89-129. doi:10.1111/1467-9531.271020

Ruef, Martin. 1999. "Social Ontology and the Dynamics of Organizational Forms: Creating Market Actors in the Healthcare Field, 1966-1994." *Social Forces* 77(4):1403-32. doi:10.2307/3005881

Rule, A., J. P. Cointet, and P. 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. doi:10.1073/pnas.1512221112. https://www.ncbi.nlm.nih.gov/pubmed/26261302.

Smith, Tammy. 2007. "Narrative Boundaries and the Dynamics of Ethnic Conflict and Conciliation." *Poetics* 35(1):22-46. doi:10.1016/j.poetic.2006.11.001

Snow, David A. and Doug McAdam. 2000 "Identity Work Processes in the Context of Social Movements: Clarifying the Identity/Movement Nexus." in *Self, Identity, and Social Movements*, edited by S. Stryker, S. J. Owens, and R. W. White. Minnesota: University of Minnesota Press, pp.41-67.

So, Richard, Hoyt Long, and Yuancheng Zhu. 2019. "Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000." *Journal of Cultural Analytics* 3(2). doi:10.22148/16.031

Spirling, Arthur and Pedro L Rodriguez. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *Journal of Politics* 84(1):101-15. doi:10.1086/715162

Stoltz, Dustin S. and Marshall A Taylor. 2019. "Concept Mover's Distance: Measuring Concept Engagement via Word Embeddings in Texts." *Journal of Computational Social Science* 2(2):293-313. doi:10.1007/s42001-019-00048-6

Stoltz, Dustin S. and Marshall A Taylor. 2021. "Cultural Cartography with Word Embeddings." *Poetics*88. doi:10.1016/j.poetic.2021.101567

Straka, Milan. 2018. "UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task." *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*: 197-207

Stuhler, Oscar. 2021. "What's in a Category? A New Approach to Discourse Role Analysis." *Poetics* 88. doi:10.1016/j.poetic.2021.101568

Tauberg, Michael. 2019. "How Smart is Your News Source?". Accessed November 22, 2021, (https://towardsdatascience.com/how-smart-is-your-news-source-1fe0c550c7d9.

Taylor, Marshall and Dustin Stoltz. 2020. "Concept Class Analysis: A Method for Identifying Cultural Schemas in Texts." *Sociological Science* 7:544-69. doi:10.15195/v7.a23

Taylor, Marshall A. and Dustin S Stoltz. 2021. "Integrating Semantic Directions with Concept Mover's Distance to Measure Binary Concept Engagement." *Journal of Computational Social Science* 4(1):231-42. doi:10.1007/s42001-020-00075-8

Tilly, Charles. 1995. *Popular Contention in Great Britain, 1758-1834*. Cambridge: Harvard University Press.

Tilly, Charles. 1997. "Parlamentarization of Popular Contention in Great Britain, 1758-1834." *Theory and Society* 26(2/3):245-73. doi:10.1023/A:1006836012345

Todorov, Tzvetan and Arnold Weinstein. 1969. "Structural Analysis of Narrative." *NOVEL: A Forum on Fiction* 3(1):70-6. doi:10.2307/1345003

Tyrkkö, Jukka. 2016. "Looking for Rhetorical Thresholds: Pronoun Frequencies in Political Speeches." *Studies in Variation, Contacts, and Change in English* 17.

Underwood, Ted. 2019. *Distant Horizons. Digital Evidence and Literary Change*. Chicago: University of Chicago Press.

Universal-Dependencies. 2020. "Universal Dependencies Framework." http://www.universaldependencies.org.

van Atteveldt, 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. doi:10.1093/pan/mpn006

van Atteveldt, Wouter, Tamir Sheafer, Shaul R. Shenhav, and Yair Fogel-Dror. 2017. "Clause Analysis: Using Syntactic Information to Automatically Extract Source, Subject, and Predicate from Texts with an Application to the 2008–2009 Gaza War." *Political Analysis* 25(2):207-22.

van Dijk, Teun A. 1972. *Some Aspects of Text Grammars. A Study in Theoretical Linguistics and Poetics*. The Hague/Paris: Mouton.

von Wiese, Leopold. 1965. *Die Philosophie der persönlichen Fürwörter*. Tübingen: Mohr.

Wada, Takeshi. 2004. "Event Analysis of Claim Making in Mexico: How are Social Protests Transformed into Political Protests?" *Mobilization* 9(3):241-57. doi:10.17813/maiq.9.3.7wx2pt66130718v3

Wagner-Pacifici, Robin. 2010. "Theorizing the Restlessness of Events." *American Journal of Sociology* 115(5);1351-1386.

Welbers, Kasper, Wouter van Atteveldt, and Jan Kleinnijenhuis. 2021. "Extracting Semantic Relations Using Syntax." *Computational Communication Research* 3(2):1-16. doi:10.5117/CCR2021.2.003.WELB

Wickham, Hadley. 2021. "babynames: US Baby Names 1880-2017." https://cran.r-project.org/web/packages/babynames/index.html.

Wodak, Ruth, Rudolf de Cillia, Martin Reisigl, and Karin Liebhart. 2009. *The Discursive Construction of National Identity*. Edinburgh: Edinburgh University Press.

Yung, Vincent. 2021. "A Visual Approach to Interpreting the Career of the Network Metaphor." *Poetics* 88. doi:10.1016/j.poetic.2021.101566

Author Biography

Oscar Stuhler is a PhD student in the Department of Sociology at New York University. He studies representations of social structures in discourse.