Chapter 28
Text as data: An overview[1]

Kenneth Benoit[2]

Version: July 16, 2019

[1] Written for *The SAGE Handbook of Research Methods in Political Science and International Relations*. I am deeply grateful for the patience shown by the editors of this volume and for their invaluable feedback on earlier versions. For feedback and comments I also thank Audrey Alejandro, Elliott Ash, April Biccum, Dan Hopkins, Cheryl Schonhardt-Bailey, Michael Laver, Stephan Ludwig, Ludovic Rheault, Arthur Spirling, and John Wilkerson. The code required to produce all of the analysis in this chapter can be viewed at `https://gist.github.com/kbenoit/4593ca1deeb4d890077f3b12ba468888`.

[2] Kenneth Benoit is Professor of Computational Social Science, Department of Methodology, London School of Economics and Political Science and Professor Part-Time, School of Politics and International Relations, Australian National University. E-mail: kbenoit@lse.ac.uk.

1 Overview

When it comes to textual data, the fields of political science and international relations face a genuine embarrassment of riches. Never before has so much text been so readily available on such a wide variety of topics that concern our discipline. Legislative debates, party manifestos, committee transcripts, candidate and other political speeches, lobbying documents, court opinions, laws — all are not only recorded and published today, but many in readily available form that is easily converted into structured data for systematic analysis. Where in a previous era what political actors said or wrote provided insight for political observers to form opinions about their orientations or intentions, the structured record of the texts they leave behind now provides a far more comprehensive, complete, and direct record of the implications of these otherwise unobservable states. It is no exaggeration, as Monroe and Schrodt (2009, 351) state, to consider text as "the most pervasive — and certainly the most persistent — artifact of political behavior". When processed into structured form, this textual record provides a rich source of data to fuel the study of politics. This revolution in the quantity and availability of textual data has vastly broadened the scope of questions that can be investigated empirically, as well as the range of political actors to which they can be applied.

Concurrent with textual data about politics becoming ubiquitous has been the explosion of methods for structuring and analysing this data. This wave has touched the shores of nearly all social sciences, but political science especially has been at the forefront of innovation in methodologies and applications of the analysis of text as data. This is most likely driven by a characteristic shared by some of the most important concepts in our discipline: they are fundamentally unobservable in any direct fashion, despite forming the foundation of our understanding of politics. Short of psychic powers or science fiction devices for reading minds, we will never have access to direct, physical measures of the content or intensity of such core concepts as ideology, commitment to democracy, or differing preferences or priorities for competing policies. Moreover, it is far from clear that every political actor even has such views or preferences, since many – such as political parties, coalition governments, or nation-states – are hardly singular actors. Even singular actors may be unaware or self-deceptive about their intentions. Behaviour provides insights into these inner states, but what political actors *say*, more than the behaviour they exhibit, provides evidence of their true inner states.

For those facing the jungle of available textual data, navigating the thicket of different approaches and methodologies for making sense of this data can be no less challenging. Widely available computational tools combined with methods from machine learning allow unprecedented insight to be drawn from textual data, but understanding and selecting from these tools and methods can be daunting. Many recent authors have surveyed the range of methodologies and their applications (e.g. Wilkerson and Casas, 2017; Lucas et al., 2015; Slapin and Proksch, 2014; Grimmer and Stewart, 2013). Rather than retrace or duplicate the efforts of my expert colleagues, here I take a slightly different tack, focusing primarily on an overview of treating text "as data" and then exploring the full implications of this approach. This involves clearly defining what it means to treat text as data, and contrasting it to other approaches to studying text. Comparing the analysis of text as data in the study of politics and international relations to the analysis of text as text, I place different approaches along a continuum of automation and compare the different research objectives that these methods serve. I outline stages of the analysis of text as data, and identify some of the practical challenges commonly faced at each stage. Looking ahead, I also identify some challenges that the field faces moving forward, and how we might meet them in order to better turn world of language in which we exist every day, into structured, useful data from which we can draw insights and inferences for political science.

2 Text, Data, and "Text as Data"

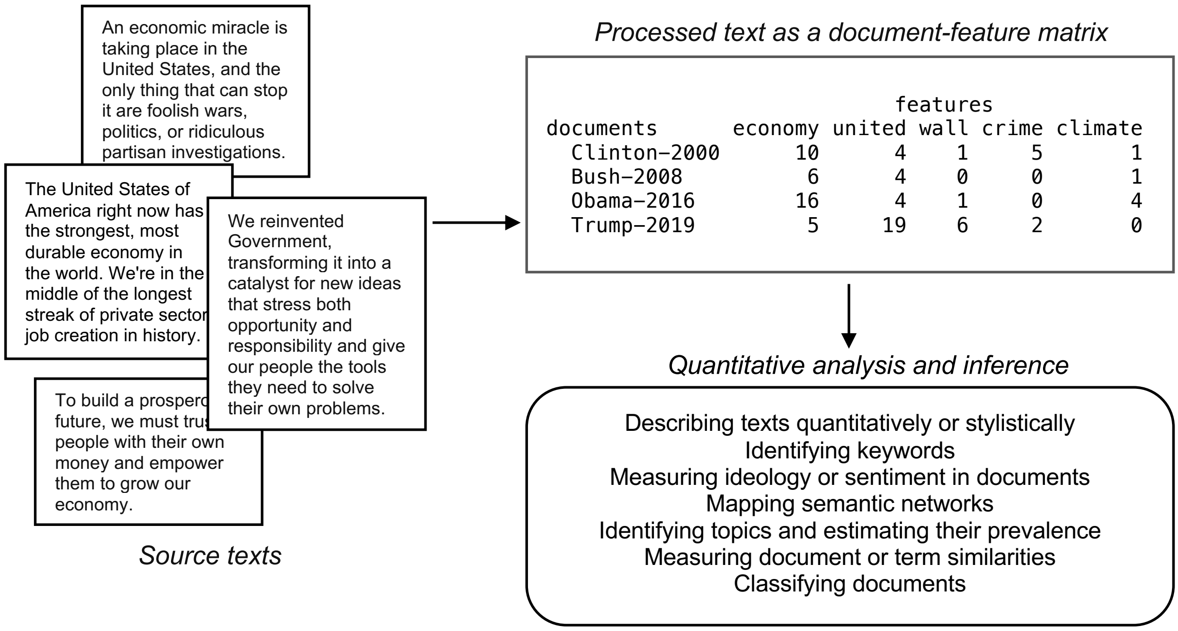
2.1 Text as text, versus text as data

Text has always formed the source material for political analysis, and even today students of politics often read political documents written thousands of years ago (Monroe and Schrodt, 2009). But for most of human history, the vast bulk of verbal communication in politics (as well as in every other domain) went unrecorded. It is only very recently that the cost of preserving texts has dropped to a level that makes it feasible to record them, or that a large amount of verbal activity takes place on electronic platforms where the text is already encoded in a machine form that makes preserving it a simple matter of storage. Official documents such as the Congressional Record that transcribes what is said in the US legislature is now supplemented by records of email, diplomatic communications, news reports, blog posts, social media posts, public speeches, and campaign documents, among others.

There is a long tradition of analysing texts to gain information about the actors who produced them (e.g. Berelson 1952), even before computerised tools became available to facilitate even traditional methods of "content analysis," defined as the human coding of texts into researcher-defined categories (Krippendorff, 2013). In a different tradition, qualitative scholars may read critically into texts as discourses to uncover the patterns and connections of knowledge and power in the social structures that produced the texts (e.g. Foucault 1972; Fairclough 2001; see van Dijk 1997 for an overview). Such data have always formed the empirical grist for the analytical mill of political science, but only in the last two decades has the approach begun to shift when it comes to treating text as something not to be read, digested, and summarised, but rather as inputs to more automated methods where the text is treated as data to be processed and analysed using the tools of quantitative analysis, even without necessarily being read at all.

The very point of text is to communicate something, so in a sense all forms of text contain information that could be treated as a form of *data*. Texts are therefore always informative in some way (even when we do not understand how). The primary objective of verbal activity, however, is not to record information, but to *communicate*: to transmit an idea, an instruction, a query, and so on. We can record it and treat it as data, but the purpose of formulating our ideas or thoughts into words and sentences is primarily communication, not the recording our ideas or thoughts as a form of data. Most data is like this: the activity which it characterizes is quite different from the data itself. In economics, for instance, it may be the economic transactions (exchanging goods or services using a medium of value) that we want to characterize, and the data is an abstraction of these transactions in some aggregated form that helps us to make sense of transactions using standardized measures. Through agreeing upon the relevant features to abstract, we can record and thus analyse human activities such as manufacturing, services, or agriculture. The process of abstracting features of textual data from the acts of communication follows this same process, with one key difference: Because raw text can speak to us directly through the language in which it is recorded, text does not first require processing or abstraction in order to be analyzed. My argument here, however, is that this process of feature abstraction is the distinguishing ingredient of the approach to treating text as data, rather than analyzing it directly as text.

Text is often referred to as “unstructured data”, because it is a (literally) literal recording of verbal activity, which is structured not for the purposes of serving as any form of data but rather structured according to the rules of language. Because “data” means, in its simplest form, information collected for use, text starts to become data when we record it for reference or analysis, and this process always involves imposing some abstraction or structure that exist outside the text itself. Absent the imposition of this structure, the text remains *informative* — we can read it and understand (on some form) what it *means* — but it does not provide a form of *information*. Just as with numerical data, we have to move from the act itself (speaking or writing) to a transformed and structured form of representing the act in order to turn the text into useful information. This is standard practice when it comes to other forms of data, but because we cannot read and understand raw numerical data in the


Source texts
Processed text as a document-feature matrix
Quantitative analysis and inference

An economic miracle is taking place in the United States, and the only thing that can stop it are foolish wars, politics, or ridiculous partisan investigations.

The United States of America right now has the strongest, most durable economy in the world. We're in the middle of the longest streak of private sector job creation in history.

We reinvented Government, transforming it into a catalyst for new ideas that stress both opportunity and responsibility and give our people the tools they need to solve their own problems.

To build a prosperc future, we must tru people with their own money and empower them to grow our economy.

```
                                  features
documents       economy united wall crime climate
  Clinton-2000       10      4    1     5       1
  Bush-2008           6      4    0     0       1
  Obama-2016         16      4    1     0       4
  Trump-2019          5     19    6     2       0
```

Describing texts quantitatively or stylistically
Identifying keywords
Measuring ideology or sentiment in documents
Mapping semantic networks
Identifying topics and estimating their prevalence
Measuring document or term similarities
Classifying documents

Figure 1: From text to data to data analysis.

same way that we can raw text, we have not yet fully equated the two processes. No one would hesitate to transform interval data such as age or income into ordinal categories of age or income ranges. (This improves accuracy at some cost of precision, as well as lessening the potential embarrassment of some survey respondents upon being asked to divulge how old they are or how little (or much) they earn.) The essence of treating text as data is that is *always* transformed into more structured, summary, and quantitative data to make it amenable to the familiar tools of data analysis.

Figure 1 portrays this process in three simple stages: raw texts, their processing and conversion into a quantitative form, and the analysis of this quantitative form using the tools of statistical analysis and inference. (I return in detail to the steps of this process below, but it is useful at this point to identify the essential stages of this process here.) Treating texts as data means arranging it for the purpose of analysis, using a structure that probably was not part of the process that generated the data itself. This step starts with collecting it into a *corpus*, which involves defining a sample of the available texts, out of all other possible texts that might have been selected. Just as with any other research, the principles of research design govern how to choose this sample, and should be guided by the research question. What distinguishes text from *textual data*, however, is that it has been selected for a research question to begin with, rather than simply representing a more fundamental act of communication by its producer. Once selected, we then impose substantial selection and abstraction in the form of converting the selected texts into a more structured form of data. The most common form in quantitative approaches to text as data is to extract *features* in the form of selected terms and tabulate their counts by documents: the "document-feature matrix" depicted in Figure 1 (and to which we will return in more detail below). This matrix form of textual data can then be used as input into a variety of analytical methods for describing the texts, measuring or mapping the targets of interest about which they contain observable implications, or classifying them into politically interesting categories.

Quantitative text analysis thus moves textual data into the same domain as other types of quantitative data analysis, making it possible to bring to bear well-tested statistical and machine learning tools of analysis and prediction. By converting texts into a matrix format, we unlock a vast arsenal of methods from statistical analysis designed for analysing matrix-type data: the comparison of distributions, scaling and measurement models, dimensional reduction techniques and other forms of multivariate analysis, regression analysis, and machine learning for prediction or identifying patterns. Many of these approaches, furthermore, are associated with well-understood properties that can be used for generating precise probability statements, such as the likelihood that an observed sample was generated from an assumed distribution. This allows us to generate insights from text analysis with precise confidence estimates, on a scale not otherwise possible.

Ironically, generating insight from text as data is only possible once we have destroyed our ability to make sense of the texts directly. To make it useful as *data*, we had to obliterate the structure of the original text and turn its stylised and oversimplified features into a glorified spreadsheet that no reader can interpret directly, no matter how expert in linear algebra. No similar lament is issued when processing non-textual data, because the form in which it can recorded as data in the first place is already a highly stylised version of the phenomena it represents. Such data began as a numerical table that we could not interpret directly, rather than as a direct and meaningful transcription of the act it recorded, whether it consisted of demographical data, (numerical) survey responses, conflict data, roll call votes, or financial indicators. Quantitative analysis is the starting point of making sense of non-verbal data, and perhaps for these reasons has never proven controversial. With text, on the other hand, we often question what is lost in the process of extracting stylised features for the purpose of statistical analysis or machine learning, because we have a reasonable sense of what is lost in the meat grinder that turned our beautiful language into an ugly numerical matrix.

This point is so important that it bears repeating. We hardly find it strange to be unable to make sense globally of a matrix of economic indicators, which we also recognise are imperfect and incomplete representations of the economic world involving the arbitrary selection of features from this world — such as the official definition of a basket of typical goods whose prices are used for measuring inflation. There is no controversy in acknowledging that while we might be able to interpret a specific figure in one cell of dataset by matching a column called inflation and a row with other columns whose values match “Canada” and “1973q3”, to make sense of more general trends we need analytical synthesis using machines. With text, on the other hand, we cannot ignore the semantic violence to our raw material and its consequences of processing our raw text into textual data, with the necessarily imperfect and incomplete representation of the source language that this requires. Machines are stupid, yet treating text as data means letting stupid machines process and perhaps analyze our texts. Any human reader would know right away that *terror* has nothing to do with political violence in sentences such as “Ending *inflation* means freeing all Americans from the terror of runaway living costs.”[1] We can only hope that our process of abstraction into textual features is smart enough not to confuse the two concepts, since once our texts have become a document-feature matrix as portrayed Figure 1, it will be hardly more interpretable than a set of raw inflation figures. In the discussion of the choice of appropriate procedure for analysing textual data, I return to this concern in more detail. The key point is that in order to text as data rather than text as text, we must destroy the immediate interpretability of source texts but for the purpose of more systematic, larger-scale inference from their stylised features. We should recognize this process unflinchingly, but also not lose any sleep over it, because the point in analysing text as data was never to interpret the data but rather to mine it for patterns. Mining is a destructive process — just ask any mountain — and some destruction is inevitable in order to extract its valuable resources.

2.2 Latent versus manifest characteristics from textual data

In political science, we are often most interested not in the text itself, but rather in what it tells us about a more fundamental, *latent* property of the text’s creator. In the study of politics (as well as psychology), some of our important theories about political and social actors concern qualities that are unobservable through direct means. Ideology, for instance, is fundamental to the study of political competition and political preferences, but we have no direct measurement instrument for recording an individual or party’s relative preference for (for example) socially and morally liberal policies versus conservative ones. Other preferences could include being relatively for or against a specific policy, such as the repeal of the Corn Laws in Britain in 1846 (Schonhardt-Bailey, 2003); being for or against further European integration during the debate over the Laeken Convention (Benoit et al., 2005); or

[1] From Ronald Reagan’s 1981 inaugural address, see `https://www.presidency.ucsb.edu/documents/presidential-documents-archive-guidebook/inaugural-addresses`.

being for or against a no confidence motion (Laver and Benoit, 2002). These preferences exist as inner states of political actors, whether these actors are legislators, parties, delegates, or candidates, and hence cannot be directly observed. Non-verbal indicators of behaviour could also be used for inference on these quantities, but it has been shown that what political actors *say* is more sincere than other forms of behaviour, such as voting in a legislature that is subject to party discipline and may be highly strategic (Herzog and Benoit, 2015). Textual data thus may contain important information about orientations and beliefs for which nonverbal forms of behavior may serve as poor indicators. The field of psychology has also long used verbal behaviour as an observable implication of underlying states of interest, such as personality traits (e.g. Tausczik and Pennebaker, 2010). Absent enhanced interrogation techniques or mind-reading technology to discern the preferences, beliefs, intentions, biases, or personalities of political and social actors, the next best alternative is to collect and analyze data based on what they are saying or writing. The target of concern is not so much what the text contains, but what its contents reveal as data about the latent characteristics for which the text provides observable implications.

Textual data might also focus on *manifest* characteristics whose significance lies primarily in how they were communicated in the the text. Much of the field of political communication, for instance, is concerned not with the latent characteristics indicated by the texts but rather with the form and nature of the communication contained in the text itself. To take a classic example, in a well-known study of articles by other Politburo members about Stalin on the occasion of his 70th birthday, Leites, Bernaut and Garthoff (1951) were able to measure differences in groups with regard to communist ideology. In this political episode, the messages signalled not only an underlying orientation but also a degree of political maneuvering with regard to a leadership struggle following the foreseeable event of Stalin's death. The messages themselves are significant, and these could only be gleaned from the public articles authored by each Politburo member, written in the full knowledge that they would be reprinted in the party and general Soviet press and interpreted as signals by other regime actors. To take another example, if we were interested in whether a political speaker used populist or racist language, this language would be manifest directly in the text itself in the form of populist or racist terms or references, and what would matter is whether they were used, not so much what they might represent. In their study of the party political broadcasts of Belgian political parties, for instance, Jagers and Walgrave (2007) established how much more overtly populist was the language used by the extreme-right party Vlaams Blok party compared to other Belgian parties.

In practice, the quality of a characteristic observerable from text as being manifest versus latent is not always sharply differentiated. Stylistic features, for instance, might be measured as manifest quantities from the text but of interest for what they tell us about the author's more fundamental traits that led to the features' use in communication. In studies using adaptations of *readability* measures applied to political texts, for instance, we might be interested either in the latent level of political sophistication as a measure of speaker intention or speaker characteristics, as evidenced by the observed sample of texts; alternatively, we might be interested in the manifest differences in their readability levels as more direct indicators of the medium of communication. In a study of historical speeches made in the British parliament, for instance, Spirling (2016) attributes a shift to simpler language in the late 19th century to the democratising effects of extending the franchise. Using similar measures, Benoit, Munger and Spirling (2019) compared a sample of US presidential State of the Union addresses delivered on the same day, by the same president, but in both spoken and written forms to show that the spoken forms used easier language. The former study might be interested in language easiness as an indicator of a more latent characteristic about the political representation, while the latter analysis might be more focused on the manifest consequences of the medium of delivery. For many research designs using textual data, the distinction is more a question of the research objective than some intrinsic way that the textual data is structured and analysed.

2.3 What "text as data" is not

We defined "textual data" as text that has undergone selection and refinement for the purpose of more analysis, and distinguished latent from manifest characteristics of the text as the qualities about which the textual data might provide inference. While this definition is quite broad, it excludes many other forms of text analysis. It is useful then, to identify the types of textual analysis that we do not consider as involving the analysis of text as data.

In essence: The study of text that does not extract elements of the text into a systematic form — into *data* — are not treating the text as data. Interpretivist approaches that focus on what a text *means* are treating the text as content to be evaluated directly, not as source material for systematic abstractions that will be used for analysis, only following which will its significance be evaluated. This is true even when the object of concern may ultimately be far greater than the text itself, such as in critical discourse analysis, whose practitioners' concern with text is primarily with social power and its abuse, or dominance and inequality as they are sustained or undermined by the text (van Dijk, 1994, 435). While this approach shifts attention to the texts as evidence of systemic injustices, the concern is more on the ability to construct a narrative of evidence for these systemic biases to be interpreted directly, rather than on extracting features from the text as data that will then be used in some analytic procedure to produce evidence for or against these injustices. The difference is subtle, but has to do with whether the interpretation of a direct reading of the text (no matter how systematic) is the end result of inquiry, versus an analysis only of extracted features of the text using a procedure that does not involve direct interpretation (such as reading) of those features. The latter treats the text as data, while the former is more focused on the text as text, to be interpreted and analysed as text.

Treating text as data is not strictly limited to quantitative approaches. Some of the most popular methods for analysing to text as data, in fact, rely on qualitative strategies for extracting textual features. Classical *content analysis*, for instance, requires reading and understanding the text. The purpose of this qualitative strategy, however, is to use content analysis to extract features from textual data, not for analysing directly what is read and understood. In reading units of the text and annotating them with predefined labels, content analysis uses human judgment not to make sense of it directly, but instead only to apply a scheme to convert the text into data by recording category labels or ratings for each unit of text. Any analysis then operates on this data, and this analysis is typically quantitative in nature even if this only involves counting frequencies of keywords or category labels. But there is nothing to say that the process of extracting the features of the text into data needs to be either automated or statistical, and in thematic and content analytic approaches, they are neither. Most direct analysis of the text without systematically extracting its features as data — text as text — is by contrast almost always qualitative because raw text is inherently qualitative. This explains why text as data approaches are associated with quantitative analysis and interpretative approaches with qualitative analysis, but to equate them would be to obscure important qualitative elements that may exist as part of a text as data research design.

Many direct forms of analysing text as text exist. The analysis of political *rhetoric*, for instance, can be characterised as the science and art of persuasive language use through "effective or efficient speaking and writing in public" (Reisigl, 2008, 96). It involves a form of discourse analysis of the text, especially with respect to the use of tropes, symbols, allegories, metaphors, and allusions. The study of anaphora in Martin Luther King's "I Have a Dream" speech (the repetition of "Now is the time..." at the beginning of sentences), for instance, involves analysing the form of its language directly, not abstracting it into data that will only then be analysed. When elements of the speech are extracted systematically into features, however, and these features subject to an analytic procedure whose interpretation can be used as an *indicator* of rhetorical quality, then the same input text *has* been treated as data.[2] This involves an act of literary brutality — the disassembly and matrix decomposition of one of the most moving speeches in US political history — but it allows us to compare Martin Luther King's speech to other pieces of political oratory on a large scale and on a common methodological footing, in way that would have been infeasible through direct interpretation.[3]

Finally, it is worth mentioning how the rapidly expanding field of natural language processing (NLP) from computer science fits within the boundaries of the text as data definition. Most computer scientists are puzzled by our use of the label, as if treating text as a form of data using quantitative tools were something new or special. This is because computer scientists' approaches *always* involve some form of automated extraction of textual features and the the processing or analysis of these using algorithmic and mathematical methods. The difference between many applications in NLP and the uses of textual data in political science lies not in whether the text is treated as data, bur rather in the purposes for which this data is used. Computer scientists are frequently concerned with engineering challenges, such as categorising structure and syntax in language, classifying or summarising documents, mapping semantic spaces, machine transition, speech recognition, voiceprint authentication, and so on. All of these are driven by textual data, but for objectives very different from the political scientist's goal of making inferences about politics. Much research in NLP concerns the use of (big) textual data to make inference about patterns in natural language. Text for political scientists, by contrast, is just one more type of informative behaviour about politics, not something whose innate properties interest us in their own right. The main advantage and objective of analysing text as data in political science is to make inferences about the same phenomena that we have long studied using non-textual data.

The key dividing line, then, involves whether the analytic procedure — whether this is interpretation, critical discourse analysis, rhetorical analysis, frequency analysis, or statistical analysis — is applied to directly to the text, or whether some intermediate step is applied to the text to extract its salient features which only then are analyzed for insight. Within this broad definition, there are many forms this can take, and in what follows I contrast these along a continuum of automation and research objective, or what I call the target of concern.

[2] Exactly such an analysis has been applied by Nick Beauchamp to the "I Have a Dream" speech. Beauchamp's "Plot Mapper" algorithm segments the text into sequential chunks, creates a chunk-term count matrix, computes the principal components of this matrix, standardises the resulting scores, and plots the first two dimensions to show the rhetorical arc of a speech. See `http://www.nickbeauchamp.com/projects/plotmapper.php`.

[3] For another cringeworthy example of procedural barbarity committed against a great political text, see Peter Norvig's "The Gettysburg Powerpoint Presentation", `https://norvig.com/Gettysburg/`.

Table 1: A map of approaches to the analysis of political text.

Approach	Method	Target of concern
Literary	Discourse analysis	Meaning Rhetoric Power relations Hidden biases Rhetoric Symbolism
Qualitative	Thematic analysis Content analysis	Topics Positions Affect Authorship Intent Similarity Events
Hybrid quantitative	Dictionary analysis	
Purely quantitative	Statistical summary Machine learning	

3 Varieties of Text Analysis

We can distinguish three main variants of text analysis, differing in whether they treat the text as information to be analysed directly versus whether they treat the text as source of data to be systematically extracted and analysed. Their situation along the continuum of text as text versus text as data can be contrasted on the basis of the degree of automation and on the target of concern to which the analysis is applied. This is a vast oversimplification, of course, but serves to contrast the essential differences between approaches.

3.1 Literary analysis

The first area is the one furthest from the approach of treating text "as data" described here: literary analysis. This approach is aimed not at treating the text as an observable implication of some underlying target of interest, but rather as the target of interest itself: text as text. In extreme forms, this may treat the text as the sole object of interest, holding the characteristics or intention of the author of the text as irrelevant. This was the view of the "New Criticism" school of literature theory advanced by Wimsatt and Beardsley (1946) in their influential essay "The Intentional Fallacy", which argued against reading into author intentions or experiences, and advocated instead focusing exclusively on the text itself.

A great many other schools of literary thought exist, of course, including postmodernist approaches that do just the opposite of avoiding reading beyond the texts, and instead examine them critically as situated in their social context. What do the texts reveal about the structures of power in a social system, especially in relation to marginalised individuals or groups? Critical discourse analysis is concerned less (or not at all) with description of the text or inference from data extracted from the text, but rather with features underlying the system in which the text occurred, even if its analysis takes place through analyzing the text as text. A critical discourse of president's speeches, for example, could focus on its commands and threats and how these are aimed at "managing the minds of others through a manipulation of their beliefs" (van Dijk 1994, 435; for an example in this context see Chilton 2017). Treating presidential speeches as data, by contrast, could consist of a computerised analysis of the words used to contrast sentiment across time or to compare different individuals (e.g. Liu and Lei, 2018). The former is interested in the text as evidence for a philosophical and normative critique of power, while the latter is concerned with supplying more empirical data on the ability to describe and compare the preferences or styles of political actors in the context of open-ended scientific propositions. Discourse analysis may be very systematic, and indeed this was a key contribution of Fairclough (2001) who developed a sophisticated methodology for mapping three distinct dimensions of discourse onto one another. The key point here concerns the role of the text in the analysis, whether it forms the end object of inquiry as a text versus whether it will be used as a source of data, with the text itself of secondary or instrumental value.

3.2 Qualitative text analysis

What I have labelled as *qualitative* approaches to the analysis of political text are distinguished from discourse analysis by focusing not on what the texts mean, either about the authors, their attempts to influence the audience, or to shore up or wear down the structures of the social system, but instead to gain more neutral empirical data from the texts by using qualitative means to extract their features. "Qualitative" is used here in its simplest form, to mean that the analytical tool does not involve statistical or numerical analysis and at its core, involves human judgment and decision rather than machines. These methods include *content analysis* and *thematic* analysis.

Sometimes called "qualitative content analysis", content analysis is the human annotation of textual content based on reading the texts and assigning them categories from a pre-defined scheme. Many of the most widely cited comparative political datasets are generated from content-analytic schemes of this type, such as the Manifesto Project (e.g. Budge, Robertson and Hearl, 1987; Budge et al., 2001). and the Comparative Policy Agendas Project (Walgrave and De Swert, 2007; Baumgartner, Green-Pedersen and Jones, 2008). Both employ human coders to read a text, segment that text into sentences or phrases, and apply fixed content codes to the segments using a pre-defined scheme that the coders have been trained to use.

Thematic analysis is essentially the same procedure, but involving a more iterative process whereby the annotation scheme can be refined during the process of reading and annotating the texts. These two approaches are closely related, since most content analytic schemes are developed by starting with a core idea and then are refined through a thematic process of attempting to apply it to a core set of texts. Thematic analysis resembles discourse analysis, and may even involve the same computer assisted tools for text annotation. It differs however in that both it and content analysis aim at a structured and more neutral and open-ended empirical approach to categorising, in the words of early political scientist Harold Lasswell, who says what, to whom, and to what extent (Lasswell, 1948). Qualitative text analysis in this tradition aims not at a critique of discourse, but rather as "a research technique for the objective, systematic and quantitative description of the manifest content of communication" (Berelson, 1952, 18).

Qualitative text analysis is labour intensive, but leverages our unique ability to understand raw textual data to provide the most *valid* means of generating textual data. Human judgment is the ultimate arbiter of the "validity" of any research exercise, and if human judgment can be used to generate data from text, we tend to trust this procedure more than we would the results of a machine — just as we would typically trust a bilingual human interpreter to render a correct translation more than we would Google Translate. This conclusion belies the unfortunate fact that humans are also notoriously unreliable, in the sense of not usually doing things the exact same way when confronted with same situation. (There are special psychological designations for those who do, including autism and obsessive-compulsiveness.) Two different human annotators, moreover, have naturally different perspectives, judgments, proclivities, and experiences, and these invariably cause them to apply an analytic scheme in different ways. In tests to replicate the Manifesto Project's scheme for annotating the sentences of manifestos, even among trained expert coders, Mikhaylov, Laver and Benoit (2012) found levels of inter-rater agreement and reliability so low that had the coders been oncologists, their levels of tumor misdiagnosis would have been medically and financially catastrophic. Methods exist to increase coder reliability, such as formulating explicit rules and carefully training coders, but these remain imperfect. Machine methods, by contrast, may generate results that are invalid or systematically wrong (if poorly designed), but at least they will be perfectly reliably wrong. This allows valuable and scarce human effort to focus on testing and calibrating machine-driven methods, without frustration of knowing that wrong answers might be due to random and uncontrollable factors.

3.3 Dictionary analysis

Dictionary analysis provides a very good example of a method in between qualitative content analysis and fully automated methods. The spread of computerised tools has made it possible to replace some or all of the analytic process, using machines that are perfectly reliable (but that don't know Karl Marx from Groucho Marx, much less what they are doing). One of the pioneering projects in what are known as *dictionary* approaches, the *General Inquirer* (Stone, Dunphy and Smith, 1966) arose in the late 1960s as an attempt to measure psychological qualities through texts as data, by counting words in electronic texts according to their membership in pre-defined psychological categories including positive and negative affect or "sentiment". Because the field of psychology also has the problem that many of its most important concepts are inner states that defy direct measurement, psychology has also long been concerned with the use of language as observable implications of a speaker or author's inner states, and some of the earliest and most ambitious dictionary-based projects have arisen in that field (e.g. also Martindale, 1975; Tausczik and Pennebaker, 2010).

In the "dictionary" approach to analysing text as data, a canonical concept or label (the dictionary "entry" or in the terminology I prefer, *key*) is identified with a series of patterns to which words in a text will be matched. These patterns, which I will call *values*, are usually considered equivalent instances of the dictionary key. A key labelled *posemo* (for positive emotion) might contain the values *kind*, *kindly*, and *kindn**, for instance, to match references to the emotional characteristic of "having or showing a friendly, generous, and considerate nature".[4] The last value is an example of a "glob" pattern match, where the "*" is a *wildcard* character that will match any or no additional characters up to the end of the term — for instance, the terms *kindness* and *kindnesses*. The false positives — words we detected but should not have — of *kindred* or *kindle* are excluded by these patterns, but so are the *kindliness* and its variants — what we could call "false negatives", or terms we should have detected but failed to do so.

This illustrates the key challenge with dictionary approaches: calibrating the matches to dictionary concepts in a valid fashion, using only crude fixed patterns as indicators of semantic content (meaning). The difficulty lies in constructing a text analysis dictionary so that all relevant terms are matched (no false negatives), but that no irrelevant or wrong terms are not (no false positives). The first problem is known as *specificity*, and is closely related to the machine learning performance measure known as *precision*. The second problem is known as *sensitivity*, and relates to the machine learning

[4] This example is taken from a very widely used psychological dictionary known as the *Linguistic Inquiry and Word Count* (2015 version). Tausczik and Pennebaker (2010)

concept of *recall*. Match too broad a set of terms, using for instance the pattern *kind**, and the matches attributed to positive emotion could wrongly include references to a popular electronic book reader. Match too specific a set of terms, such as *kind* only, and we would fail to match its adverbial form "kindly".

Thus far we have focused on variants distinguished by spelling, but the problem can be even more fundamental because many words spelled identically may have completely different meanings. This quality known as *polysemy*, and especially afflicts text as data approaches in English. To continue our example, *kind* may also be a noun meaning "a group of people or things having similar characteristics", such as "more than one *kind* of text analysis", or an adverb meaning "to some extent", such as "dictionary calibration can get *kind* of tricky". To illustrate, I used a part of speech tagger and some frequency analysis to distinguish the different meanings from the State of the Union corpus of presidential addresses. Of the 318 uses of *kind*, nearly 95% were the noun form while only 4% referred to the adjective denoting positive emotion (three more matches were to the "kind of" usage). It is unlikely that human annotators would confuse the noun form with the adjective indicating a positive emotion, because their qualitative data processing instruments — their human brain, with its higher thoughts structured by language itself – would instantly recognise the difference. Human judgment is also inconsistent, however, and in some rare cases a qualitative annotator could misinterpret the word, might follow their instructions differently, or might simply make a mistake. The computer, on the other hand, while mechanistically matching all occurrences in a text of the term *kind* with the category of positive emotion, will produce 95% false positive matches by including the term's non-emotional noun form homograph, but do so with perfect consistency.

This discussion is not meant to disparage dictionary approaches, as they remain enormously popular and extremely useful, especially for characterising personality traits or analysing political sentiment. They also have the appeal of easy interpretability. While building the tools to efficiently count matches of dictionary values to words in a text might require some deft engineering, the basic idea is no more complicated than a counting exercise. Once counted, the analysis of these counts uses the same simple scales that are are applied to content analysis counts, such as the percentage of positive minus negative terms. Conceptually, dictionary matches are essentially the same as human-coded content analysis, but in a cruder, more mechanistic way. Content analysis uses human judgment to apply a set of category labels to units of texts using human judgment after reading the text. Dictionary analysis replaces this with automated pattern matching to count category labels using automatic matching of the values defined as matches for those labels with words or phrases in the text. Both methods result in the construction of a matrix of texts by category counts, and from that point onward, the methods of analysis are identical. The target of concern of both approaches, as well as of the purely quantitative approaches discussed below, may be topics, positions, intentions, or affective orientations, or even simple events, depending on the coding or dictionary scheme applied and the methods by which the quantitative matrix is scaled.

Dictionary methods are listed as "hybrid" approaches because while they involve machines to match the dictionary patterns to the texts, constructing the set of matches in the dictionary is entirely a matter for human judgment. At some point, some human analyst made the judgment call to put *kind* as a match for "positive emotion" rather than (for instance) *kind**, but decided not to include (or simply overlooked) *altruistic* and *magnanimous*. Depending on the educational level of the texts to which this dictionary is applied, it will fail to various degrees to detect these more complicated, excluded synonyms. Many dictionaries exist that have been used successfully in many highly cited publications, but this is no guarantee that they will work for any untested application. In their attempt to use the venerable Harvard Psychosociological Dictionary to detect negative sentiment in the annual reports of public corporations, Loughran and McDonald (2011) for instance found that almost three-fourths of their matches to the unadjusted Harvard dictionary category of 2,010 negative words were typically not negative in a financial context: words such as *tax*, *cost*, *liability*, and *vice*. Only through a careful, qualitative process of inspection of the word matches in context were they able to make significant changes to the dictionary in a way that fit their context, before trusting the validity of results after turning the machine loose to apply their dictionary to a corpus of 50,000 corporate reports.

3.4 Statistical summaries

Statistical summary methods are essentially quantitative summaries of texts to describe their characteristics on some indicator, and may use statistical methods based on sampling theory for comparison. The simplest such measures identify the most commonly occurring words, and summarize these as frequency distributions. More sophisticated methods compare the differential occurrences of words across texts or partitions of a corpus, using statistical association measures, to identify the words that belong primarily to sub-groups such as those predominantly associated with male- versus female-authored documents, or Democratic versus Republican speeches.

Measures of similarity and distance are also common in characterizing the relationships between documents or terms. By treating each document as a vector of term occurrences — or conversely, each feature as a vector of document occurrences — similarity and distance measures allow two documents (or features) to be compared using bivariate measures such as the widely cosine similarity measure or Pearson’s correlation coefficient, or one of the many distance measures such as the Euclidean or Jaccard distance. Such measures form the backbone of the field of information retrieval but also allow comparisons between documents (and authors) that might have a more substantive political interpretation, such as ideological proximity. When generalized by comparing “local alignments” of word sequences, similarity measures also form the basis of *text reuse* methods, which have been used to study the origins of legislation (Wilkerson, Smith and Stramp, 2015) and the influence of interest groups (A and Kashin, 2015).

Other quantitative summary measures of documents are designed to characterize specific qualities of texts such as their *readability* — of which the Flesch (1948) reading ease measure is probably the best-known — or *lexical diversity*, designed to measure vocabularity diversity across a text. Sentiment analysis may also take the form of a statistical summary, such as a simple ratio for each text comparing counts of positive to negative sentiment words. Many other statistical summary measures are possible, and these may directly provide interesting quantities of interest from the text as data, depending on the precise research question. While such indexes are traditionally not associated with stochastic distributions, it is possible to compute confidence intervals for these based on bootstrapping (Benoit, Munger and Spirling, 2019) or averaging measures computed across moving windows of fixed text lengths (Covington and McFall, 2010), in order to judge statistically whether an observed difference between texts is significant.

3.5 Supervised machine learning

In the final step along the continuum of automation versus human judgment, we have machine learning methods that require no human analytical component, and are performed entirely by machine. Of course, human judgement is still required to select the texts for input or for training the machine, but this involves little more than a choice of which texts to input into the automated process. In purely quantitative approaches to text as data, there are choices about the selection and processing of inputs to be made, but not in the design of the instrument for processing or analyzing the data in the way that dictionary approaches involve.

In purely quantitive approaches, it may not only be unnecessary to read the texts being analysed, but also unnecessary for it to be *possible* to read them. Provided we have the means to segment the texts (usually, into words), then unsupervised approaches to scaling positions, identifying topics, or clustering texts can happen without any knowledge of the language itself. Even supervised methods do not require the training texts to be read (although it is reassuring and preferable!), provided that we are confident that the texts chosen are good representatives of the extremes of the positions we would like to scale. For unsupervised scaling methods, no reading knowledge is required, *if we are confident that the texts are primarily about differences over well-defined issues*. For topic modelling, not even that is required. Of course, validation is crucial if we are to trust the results of automated methods, and this almost always involves human judgment and interpretation. Having skipped human judgment as part of the analytical process, in other words, we bring back our judgment at the conclusion of the process in order to make sense of the results. If our better judgment indicates that something is askance, we may choose to adjust the machine or its inputs and repeat the process until we get improved results. This cycle is often repeated several times, perhaps with different model parameters, for such tasks as classification, topic models (especially for choosing the number of topics), document selection for unsupervised scaling, or more fine-grained adjustment such as feature selection. The choice of machine and its settings are important, but the ability to make sense of the words has become unimportant. This approach works with any language, because in stark contrast to the literary methods in which the meaning of words is the target of concern, "it treats words simply as data rather than requiring any knowledge of their meaning as used in the text" (Laver, Benoit and Garry, 2003, 312). In fully automated and quantitative approaches, the words are merely signals to help us interpret the political phenomena that gave rise to them, much as astronomers interpret minute variations in light wavelengths to measure more the fundamental targets of concern affected them, such as planetary sizes and orbits.

Supervised machine learning is based on the idea that a procedure will "learn" from texts about which the analyst declares some external knowledge, and the results of this learning are then mapped onto texts about which the analyst lacks this knowledge. The objective is inference or prediction about the unknown texts, in the same domain as the input knowledge. Classifiers based on supervised examples start with a *training set* of texts with some known label, such as *positive* or *negative*, and learn from the patterns of word (feature) frequencies in the texts to associate orientations of each word. These orientations are used for projections onto a *test set* of documents whose label is unknown, based on some aggregation of the learned word feature orientations given their observed frequencies in the unknown documents. While they perform this learning and prediction in different ways, this basic process is common to classifiers such as Naive Bayes (Pang, Lee and Vaithyanathan, 2002), SVMs (Joachims, 1999), random forests (Fang and Zhan, 2015), neural networks (Lai et al., 2015), and regression-based models (e.g. Taddy, 2013).

When applied to estimating quantities on a continuously output scale rather than class prediction, supervised machine learning techniques may be adapted for *scaling* a dimension that is “known” by virtue of the training examples used to fit the model. This is the approach of the *Wordscores* model (Laver, Benoit and Garry, 2003) that has been widely used in political science to scale ideology, as well its more modern descendant of *class affinity* scaling (Perry and Benoit, 2018). Both methods learn word associations with two contrasting “reference” classes and then combine these with word frequencies in texts whose positions are unknown, in order to estimate their positions with respect to the reference classes.

Supervised *scaling* differs from supervised *classification* in that scaling aims to estimate a position on a latent dimension, while classification aims to estimate a text’s membership in a latent class. The two tasks differ in how greedily they demand input data in the form of more features and additional documents. Typically, classification tasks can be improved by adding more training data, and some methods such as convolutional neural networks (Lai et al., 2015) require very large training sets. To minimise classification error, we may not care what features are used; as long as the model is not overfit the primary goal is simply to correctly predict a class *label*. Scaling, on the other hand, is designed to isolate a specific dimension on which texts are to be compared and provide a point estimate of this quantity, on some continuous scale. Validating this quantity is much harder than in class prediction, and typically involves comparison to external measures to establish its validity.

Unlike classification tasks where accuracy is the core objective, supervised scaling approaches have been shown capable of producing valid and robust scale estimates even with relatively small training corpora (see Klemmensen, Hobolt and Hansen, 2007; Baek, Cappella and Bindman, 2011). The key in scaling applications is more one of the quality of training texts — making sure they contain good textual representations of the opposing poles of a dimensional extreme (Laver, Benoit and Garry, 2003, 330) — than of their quantity. For scaling applications, training texts only need to contain strong examples of lexical usage that will differentiate the dimensional extremes, such as strong “conservative” language in one set, contrasting with strong “liberal” language in another, using the same lexicon that will be used in the out-of-sample (or in the words of Laver, Benoit and Garry 2003, “virgin”) texts. One advantage of not being concerned with classification performance is that scaling is robust to irrelevant text in the virgin documents. Training texts that contain language about two extremes of environmental policy, for instance, are unlikely to contain words about health care. Scaling an unknown text using a model fitted to these environmental texts will therefore scale only the terms related to (and hence only the dimension of) environmental policy, even if the document being scaled contained out-of-domain text related to health care. For unsupervised methods, by contrast, irrelevant text will seriously affect the scaled results.

Most political scientists are interested more in measurement and scaling than in classification, which is typically of only instrumental value in estimating or augmenting a dataset for additional testing. In their study of echo chambers on the Twitter social media platform, for instance, Colleoni, Rozza and Arvidsson (2014) used supervised learning trained on Tweets from around 10,000 users known to be Republican or Democrat, to predict the party affiliation of an additional 20 million users. They used the supervised classifier to augment their dataset of social media with a label of party affiliation, which is not part of the social media data but which was nonetheless central to their ability to measure partisan homophily in communication networks. Classification in social science is generally more useful in augmenting data rather than representing an interesting finding in its own right. While classifying a legislator's party affiliation might be an interesting engineering challenge for a computer scientist, this yields no new insight for a political scientist, as this information is already known (which does not mean that it has not been done, however: see Yu, Kaufmann and Diermeier 2008). Estimating the sincere political preference of a legislator whose vote is uninformative because of party discipline, by comparison, is typically of great interest in political science.

3.6 Unsupervised machine learning

Unsupervised learning approaches are similar to supervised methods, with one key difference: there is no separate learning step associated with inputs in the form of known classes or policy extremes (if scaling). Instead, differences in textual features are used to infer characteristics of the texts and their features, and these characteristics are interpreted in substantive terms based on their content or based on their correlation with external knowledge. A grouping might be labelled based on its association with different political party affiliations of the input documents, for instance, even though the party affiliations did not form part of the learning input. Examples of unsupervised methods associated with text are clustering applications, such as *k*-means clustering (see Grimmer and Stewart, 2013, 6.1), designed to produce a clusters of documents into *k* groups in way that maximises the differences between groups and minimises the differences within them. These groups are not labelled, and so must be interpreted *ex post* based on a reading of their content or the association of the documents with some known external categories. Because this is primarily a utility device for learning groups, it has few applications in political science outside of a data augmentation tool, although it has been used as a topic discovery tool in some applications, such as Sanders, Lisi and Schonhardt-Bailey (2017) who used clustering as one method to identify economic policy topics from UK select committee oversight hearings.

An unsupervised learning method that has received wide application is the latent Dirichlet allocation (LDA) *topic model* Blei, Ng and Jordan (2003). Topic models provide a relatively simple, parametric model describing the relationship between clusters of co-occurring words representing "topics" and their relationship to documents which contain them in relative proportions. By estimating the parameters of this model, it is possible to recover these topics (and the words that they comprise) and to estimate the degree to which documents pertain to each topic. The estimated topics are unlabelled, so a human must assign these labels by interpreting the content of the words most highly associated with each topic, perhaps assisted by contextual information. No human input is required to fit the topics besides a document-feature matrix, with one critical exception: the number of topics must be decided in advance. In fitting and interpreting topic models, therefore, a core concern is choosing the "correct" number of topics. There are statistical measures (such as *perplexity*, a measure based on comparing model likelihoods; or *topic coherence*, based on maximising the typical pairwise similarity of terms in a topic) but a better measure is often the interpretability of the topics. In practice the precise choice of topics contains a degree of arbitrariness, and often to recover interpretable topics, some extra ones are also generated that are not readily interpretable.[5]

Political scientists have made widespread use of topic models and their variants, including some novel methodological innovations driven by the specific demands of political research problems. Quinn et al. (2010) shifts the *mixed membership* model of topics within documents to a time unit (days in the U.S. Senate) and estimates the membership of texts with each time unit (speeches made on that day) as representing a single topic. Combined with some prior information, this model produced estimates of the daily attention to distinct political topics, to track what the Senate was talking about over a long time series. Another variation is Grimmer (2010)'s "expressed agenda model", which measures the attention paid to specific issues in Senators' press releases, based on the idea that each Senator represents a mixture of topics and will express these through individual press releases. Another innovation for which political scientists should be proud is the *structural topic model* (Roberts et al., 2014) which introduces the ability to add covariates in the form of categorical explanatory variables to explain topic prevalence. In their paper introducing this method, Roberts et al. (2014) apply it to open-ended survey responses on immigration questions to show differences in the estimated proportions of topics pertaining to fear of immigration, given the treatment effect of a survey experiment and conditioning variables related to whether a respondent identified with the Democratic

[5] For a deeper general discussion of these issues, see Steyvers and Griffiths (2007).

or Republican party. In each of these innovations, political scientists have adapted a text mining method to specific uses enabling inference about differences between time periods, individuals, or treatments, turning topic models from an exploratory tool into a method for testing systematic propositions that might relate to fundamental political characteristics of interest.

Another unsupervised method not only widely applied but also developed by political scientists is the unsupervised *wordfish* scaling model Slapin and Proksch (2008). This model assumes that observed counts in a document-feature matrix are generated by a Poisson model combining a word effect with a parameter representing a position on a latent dimension, conditioned by both document and feature fixed effects. It produces estimates of a document's latent position, which can be interpreted as left-right ideology (Slapin and Proksch, 2008), preference over environmental policy (Klüver, 2009), support or opposition to austerity in budgeting (Lowe and Benoit, 2013), or preferences for the level of European integration (Proksch and Slapin, 2010). One limitation of this model, however, is that it permits estimation on only a single dimension (although other dimensional estimates using similar methods are possible, as Monroe and Maeda 2004 and Däubler and Benoit 2018 have demonstrated). In a detailed comparison of scaling model estimates to ratings of the same texts by human coders, Lowe and Benoit (2013) showed that an anti-system party appeared wrongly (according to the human raters) in the middle of the the scale of support or opposition to the budget, because of its differences on a dimension of politics not captured in a single government-opposition divide. Because they are anchored according to extremes identified by the user, supervised scaling methods such as Wordscores can extract different positional estimates from the same texts (provided the training inputs for these texts were different). Unsupervised scaling, however, will always produce only one set of estimates for the same texts. When an analyst wants to estimate multiple dimensions, the only recourse is to input different texts. When Slapin and Proksch (2008) used their scaling method to estimate policy positions from German party manifestos on three separate dimensions of economic, social, and foreign policy, they first had to segment each manifesto into new documents containing only text relating to these themes (which required reading the texts, in German, and then manually splitting them). To control the outputs from unsupervised methods, one must control the inputs.

Poisson scaling (e.g. the wordfish method) is very similar to older methods to project document positions onto a low-dimensional space, after singular value decomposition (and some additional transformation) of the high-dimensional document-feature matrix. Such older methods include *correspondence analysis* (CA Greenacre, 2017) and *latent semantic analysis* (LSA Landauer, Foltz and Laham, 1998), both forms of metric scaling that can be used to represent documents in multiple dimensions (although LSA is more commonly used as a tool in information retrieval). These lack some advantages of parametric approaches, such as the ability to estimate uncertainty using outputs from the estimation of statistical parameters, but have nonetheless seen some application in political science because of their ease of computation and ability to scale multiple dimensions (e.g. Schonhardt-Bailey, 2008).

Because unsupervised scaling methods take a matrix as input, and this matrix might just as easily have been transposed (swapping documents for features), these methods also permit the measurement and scaling of word features as well as documents. The metric scaling from CA, for instance, allows words to be located in the same dimensional spaces as documents (see Schonhardt-Bailey, 2008, for instance). Wordfish scaling also allows us to estimate the policy weight and direction, similar to a discrimination parameter from an item-response theory model, for each feature. When features are policy categories, this can provide information of substantive interest in its own right, such as how different policies form the left-right "super-dimension" and how these might differ across different political contexts (Däubler and Benoit, 2018).

3.7 Distributional semantic models and "word embeddings"

A final exciting area deserving mention are text as data approaches based on matrices of observed words but weighted by their "word vectors", estimated from fitting a *distributional semantic model* (DSM) to a large corpus of text, often a corpus separate from the text to be analyzed as data in a given application. The notion of distributional semantics was famously articulated by the linguist John Firth, who stated that "You shall know a word by the company it keeps" (Firth, 1957, 11). Using a "continuous bag of words model" to estimate word co-occurrences within a specified context (for instance a window of five words before and after), models can be fit to estimate a vector of real-valued scores for each word representing their locations in a multi-dimensional semantic space. Known collectively as *word embedding* models, such methods provide a way to connect words according to their usages in a way that offers potentially vast improvements on the context-blind "bag of words" approach.

Relatively new methods for fitting DSMs include the *word2vec* model (Mikolov et al., 2013) that that uses a "skip-gram" neural network model to estimate the probability that a word is "close" to another given word, the *GloVe* model ("global vectors of words", Pennington, Socher and Manning, 2014) that predicts surrounding words using a form of dynamic logistic regression, and the ELMo model ("Embeddings from Language Models" Peters et al., 2018). All of these methods are widely available in open-source software implementations.

Word embedding models are usually not thought of as method on their own for analyzing text as data, but rather as extremely useful complements to representations of text as data based on word counts, and have been shown to greatly improve performance for applications such as text classification, sentiment analysis, clustering or comparing documents based on their similarities, or document summarization. Estimated from a user's own corpus, furthermore, word embeddings allow the direct exploration of semantic relations in their own verbal context, to determine the associations of terms far more closely related to their meanings than can simple clustering or similarity measures from bag-of-words count vectors.

For users that cannot fit local embedding models to a corpus, pre-trained word vectors are available that have been estimated from large corpora, such as that trained on six billion tokens from Wikipedia and the "Gigiword" corpus (Pennington, Socher and Manning, 2014). This allows a researcher to represent his or her texts not just from the corpus at hand, but also augmented with quantitative measures of the words' semantic representations fitted from other contexts. This provides an interesting twist on the discourse analytic notion of *intertextuality* (e.g. Fairclough, 1992), a process wherin the meaning of one text shapes the meaning of another. Incorporating semantic representations fitted from large corpora into the analysis of text is a literal recipe for reinforcing the pre-dominant social relations of power as expressed in language, a problem that has not gone unnoticed. Bolukbasi et al. (2016) and Caliskan, Bryson and Narayanan (2017) show that word embeddings encode societal stereotypes about gender roles and occupations, for instance that engineers tend to be men and that nurses are typically women. Data and quantification do not make our textual analyses neutral, and we should be aware of this especially when incorporating semantic context into text as data approaches.

4 The Stages of Analysing Text as Data

We have described the essence of the approach of treating text as data as involving the extraction and analysis of features from text to be treated as data, either about the manifest characteristics of the text itself or of latent characteristics for which the text provides observable implications. In this section, I describe this process in more detail, outlining the steps involved and critically examining the key choices and issues faced in each stage.

Selecting texts: Defining the corpus. A "corpus" is the term used in text analysis to refer to the set of documents to be analysed, and that have been selected for a specific purpose. Just as with any other

Table 2: Stages in analyzing text as data.

1. Selecting texts and defining the corpus.
2. Converting of texts into a common electronic format.
3. Defining documents and choosing the unit of analysis.
4. Defining and refining features.
5. Converting of textual features into a quantitative matrix.
6. Analyzing the (matrix) data using an appropriate statistical procedure.
7. Interpreting and reporting the results.

research design, research built on textual data begins with the analyst identifying the corpus of texts relevant to the research question of interest and gathering these texts into a collection for analysis. Texts are generally distinguished from one another by attributes relating to the author or speaker of the text, perhaps also separated by time, topic, or act. A year of articles about the economy from *The New York Times*, for instance, could form a sample for analysis, where the unit is an article. A set of debates during (one of the many) votes on Brexit in the UK House of Commons could form another corpus, where the unit is a speech act (one intervention by a speaker on the floor of parliament). German-language party election manifestos from 1949 to 2017 could form a corpus, where a unit is a manifesto. A set of Supreme Court decisions from 2018 year could form a corpus, where the unit is one opinion. In each example, distinguishing external attributes, chosen by the researcher for the purpose of analysing a specific research question, are used to define the *document* distinguishing one unit of textual data from another.

In many political science applications using textual data, the “sample” of texts may, in fact, be every known text generated by the political universe for that application. In tracking the words spoken on abortion per day in the U.S. Congress, for instance, a study might examine every utterance in the Senate from 1997 to 2004. Yet even in such situations where a researcher may not face overt decisions to sample texts from a larger population that is too large to cover in its entirety, such as how many newspaper articles to select from which set of days, it is still important to be aware of selection issues that shape what sort of text becomes a recorded feature of the social system. Such “social bookkeeping” has long been noted by historians seeking texts to gain leverage on events long past, but it may also feature in many forms of available political corpora, especially spoken text in structured settings such as legislatures. Historical coalition political manifestos, for instance, are notoriously difficult to obtain because they tend to disappear once a coalition has broken down, creating a potential sample bias slanted to more stable coalitions. The key is to be aware of the mechanisms governing the generation of text, with an aim to making sure that the observable text provides representative coverage of the phenomenon that it will be used to investigate.

Some sampling choices may be motivated on practical grounds, especially resource limitations. In text as data approaches pre-dating the availability of computerised tools, it was not uncommon to suggest examining 100-word samples from a text for measuring such quantities as the readability of a text (e.g. Gunning, 1952) or taking "all the words in 16 two-page groups spread uniformly throughout the book" for a measure of lexical diversity (Herdan, 1955, 332). In the modern era, by contrast, down-sampling may be required due to access limitations or because of the sheer volume of available data. The Twitter streaming API, for instance, has an overall rate limit of one percent of all Tweets for those without access to Twitter's exclusive "firehose" of all Tweets. Even researchers who have captured the tens of millions of daily Tweets available within this rate limit may decide to work on a random sub-sample of this dataset, because of the computational and storage costs involved in trying to analyse the larger dataset.

Converting the texts into a common electronic format. This step is purely technical, involving no research design decisions, but it can nonetheless pose one of the stickiest problems in text analysis. Strictly speaking, it is not necessary to work with computers to treat texts as data. The old-school methods for computing textual readability, for instance Gunning's FOG index referenced above, or applying the complex rules from Spache (1953) to match words in a text to a list of "familiar" terms, could involve working with pen, printed text, an abacus, and a lot of coffee (possibly while working by candlelight and wearing a hairshirt). In almost all contemporary applications, however, texts are sourced and processed using computers. The problem is that there are vast differences in the formats for recording electronic texts, including Adobe's "pdf" format, which is actually a collection of different variants and versions; markup languages (such as HTML or XML); word processing formats (such as Microsoft Word, which also exists in many different versions); spreadsheet formats (such as Microsoft Excel); key-value schemes (such as JSON); or, if one is really lucky, plain text (.txt) files requiring no special handling. Even plain text files, however, can require a form of conversion, since the machine *encoding* of text has many different forms, especially in the pre-Unicode era from about 1970-2000 when the same set of 8-bit numeric values were mapped to different characters depending on the computer operating system and the national context.[6] Unicode has replaced this, by providing a

[6] The history of encoding is a long and complicated saga that most practitioners of text analysis would happily ignore. It has to do with how the original 7-bit (containing 128 characters, or 2^7) "ASCII" character set needed adaptation to new languages and symbols by adding an eighth bit. There was very little standardisation in how the resulting additional 128 characters were mapped, so that text encoded for example in Windows-1250 (for Central and East European languages) would look garbled on

single, comprehensive mapping of unique code points to every known character in the world's writing systems, present and past, including emoji and special symbols. Because Unicode is a standard, however, rather than an *encoding*, it still needs to be implemented on machines, and Unicode also covers standards for this encoding, such as UTF-8 (the most common).

Conversion of images into text is another possible headache, especially for older documents that have been scanned from print. To convert these "image-only" documents, which may exist as pdf but not contain actual text, *optical character recognition* may be needed: the conversion of images of characters into electronically encoded text. Depending on the quality of the images, this can require a great deal of manual correction and cleaning. To the human eye, there may be no essential difference between *OCR* and *0CR*, but to a computer these are completely different words. Other challenges can involve typographic ligatures (such as the "fi" in *find*) and other typographic relics such as the medial *s*, printed as *f*, which was disused by around 1800 but widely found in 19th century printing. Most OCR, however, will not recognise that *Congrefs* is the same as *Congress*.[7]

Defining documents and choosing the unit of analysis. This step differs from the selection of the corpus in that prior to extracting textual features for analysis, the unit of analysis may need further definition, through selection or sampling or by aggregating documents into larger units or splitting them into smaller ones. The attributes that differentiate source texts, in other words, may not form the ideal units for analysing the text as data. (Note that by "units", here we refer to the document units, not textual features, which are covered next.)

For example, while we might have a corpus of social media posts, these might be better aggregated over some time period, such as a day, or by user. This not only ameliorates a possible problem with overly short documents, but also focuses attention on the unit of interest. Whether this is time or a user (or speaker or other unit of authorship) will depend on the research problem. For other problems, segmenting a document into smaller units might be the answer. These could be structural units such as sentences or paragraphs, or some fixed-length chunk of tokens (segmented words). Fixed-length chunks are especially useful for sampling schemes, for instance in measuring textual characteristics using a moving average across a fixed window size of a text (e.g. **?**). Some schemes may combine these approaches, such as Labbé, Labbé and Hubert (2004, c. 209)'s analysis of Charles de Gaulle's broadcast speeches from June 1958 – April 1969: first they combined these into a single

a system using the similar, but not identical ISO-8859-2 for words like *źródło*.

[7] This also explains the apparently widespread usage in the 1700s of the "f-word": not even Google Books has been able to distinguish it from the work *suck*. See `https://books.google.com/ngrams/graph?content=fuck&year_start=1700&year_end=2000`.

"document" where each speech was in the order broadcast, and then applied a form of moving average measure of lexical diversity on segments that overlapped the original document boundaries defined by the speech dates.

Identifying units of analysis may also be done qualitatively, based on reading the texts and identifying politically relevant units of text. The best known example in political research is the "quasi sentence" that forms the unit of analysis for the long-running Comparative Manifesto Project. Quasi sentences are textual units that express a policy proposition and may be either a complete natural sentence or part of one. Because some authors may express two distinct policy statements within a single natural sentence, the use of quasi sentences supposedly permits a more valid and complete representation of the content of the textual data. The trade-off, however, is that the same human decision process that interprets the sentence structure to identify text units also causes the procedure to be unreliable and often difficult or impossible to replicate (Däubler et al., 2012). This trade-off between reliability—whether repetition of a procedure produces stable results—and validity—whether the measurement or analysis reflects the truth of what is being measured or represented by the textual data—is a recurrent theme in research involving textual data. This affects not just the identification and preparation of units for analysis but also the design of coding frames and measurement and scaling models for analysing textual data.

The ability to redefine documents in terms of the smaller textual units they contain illustrates a curious feature of textual data: that the units of analysis are defined in terms of collections of the features. If we think of this data in matrix form (such as the intermediate stage of Figure 1), then the units of analysis are represented by the rows of documents and the features as columns derived from terms — indeed, this matrix is usually called a *document-term matrix*. Since a document is just an arbitrary collection of terms, however, it means that the more we segment our document into smaller collections, the more it approaches being a features unit defined by the column dimension of the data. Conversely, grouping documents together does the opposite. Redefining the boundaries of what constitutes a "document", therefore, involves shifting data from columns into rows or vice versa. This ability to reshape our data matrix because one dimension is defined in terms of a collection of the other is unique to text analysis. We could not perform a similar reshaping operation on say, a survey dataset where by spreading an individual's observed responses across additional rows, because we cannot split an individual as a unit and because that individual is defined in terms of a sampled, physical individual, not as an arbitrary collection of survey questions.

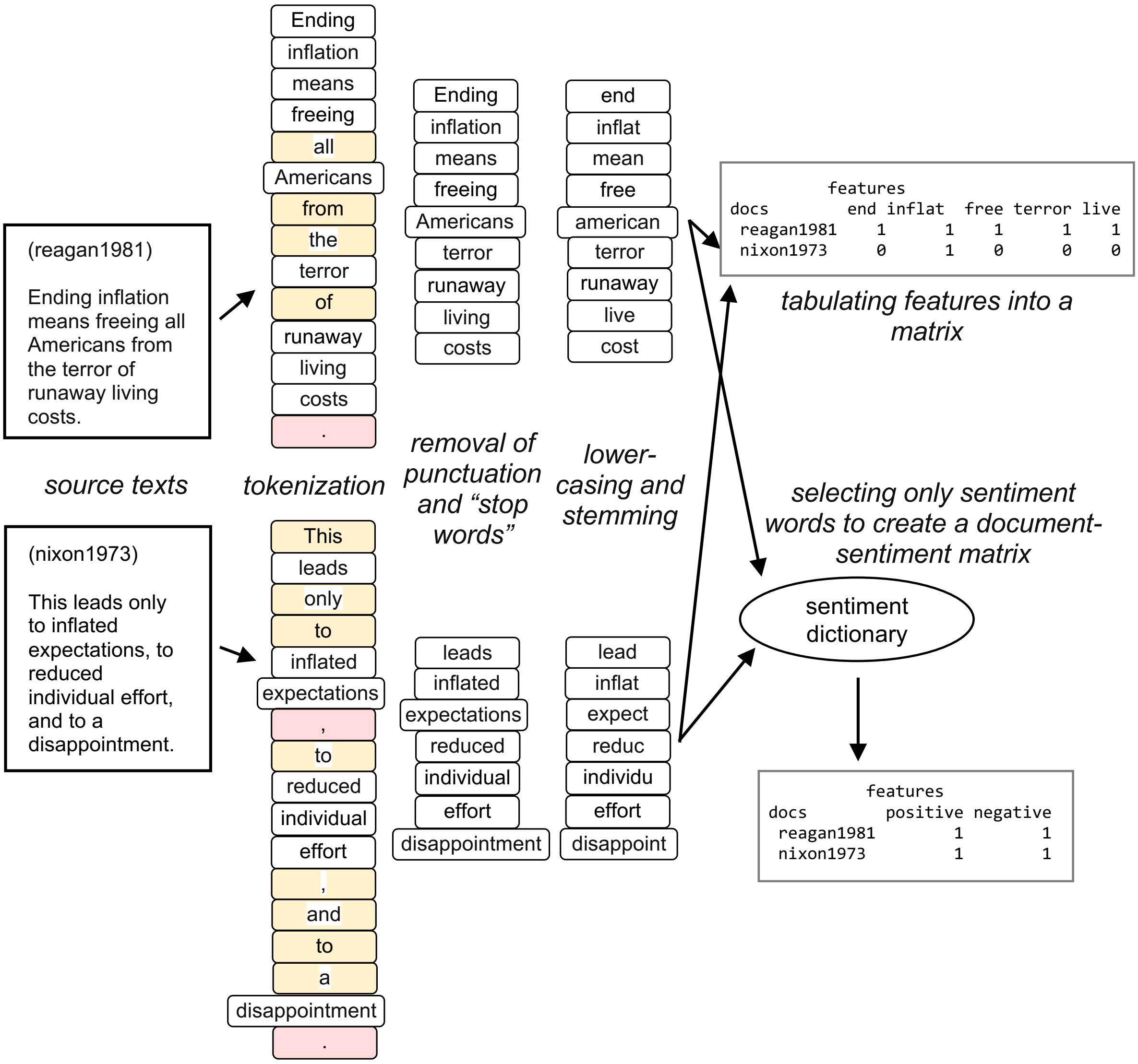
Ultimately, how we reshape our documentary units by grouping or splitting them will depend on our research question and the needs of our method for analysing the data. Knowing how the sampling procedure for the textual data selection relates to the sampling units and the units of analysis may have implications for subsequent inference, given that the units of analysis are not randomly sampled textual data, irrespective of the sampling units. Determining which are most suitable will depend on the nature of the analytical technique and the insight it is designed to yield, and sometimes the length and nature of the texts themselves.

Defining features. Features start with the basic semantic unit of text: the word. There are many forms of "words", however, and these typically undergo a process of selection and transformation before they become features of our textual dataset. Words might also simply form the basis for recording an abstraction triggered by the word, such as a dictionary key, or even a category assigned by a human annotator (in manual content analysis).

First, we should become familiar with some terms from linguistics. Words as they occur in a text are commonly known as *tokens*, so that the text "one two one two" contains four tokens. *Tokenization* is the process of splitting a text into its constituent tokens, as in the second column of Figure 2 (which includes punctuation characters as tokens). Tokenization usually happens by recognising the delimiters between words, which in most languages takes the form of a space. In more technical language, inter-word delimiters are known as *whitespace*, and include additional machine characters such as newlines, tabs, and space variants.[8] Most languages separate words by whitespace, but some major ones such as Chinese, Japanese, and Korean do not. Tokenizing these languages requires a set of rules to recognise word boundaries, usually from a listing of common word endings. Smart tokenizers will also separate punctuation characters that occur immediately following a word, such as the comma after *word* in this sentence. To introduce another term, word *types* refer to uniquely occurring words. So in our example, the four-token text contains only two word types, *one* and *two*. Comparing the rates of types and tokens forms the foundation for measures of lexical diversity (the rate of vocabulary usage), with most common such measure comparing the number of types to the number of tokens (the "type-token ratio").

For a token to become a *feature* of textual data, it typically undergoes transformation in a step often called "pre-processing" (although it should really be called processing). The most common types of token processing are *lower-casing*, which treats words as equivalent regardless of how they were capitalised; *stemming*, which reduces words to their stems, which is a cruder algorithmic means of

[8] Not Klingons, but rather the variations on the simple space character included in the Unicode "Separator, Space" category, such as `U+205F`, the "Medium Mathematical Space".



```
          features
docs         end inflat  free terror live
  reagan1981    1      1    1      1    1
  nixon1973     0      1    0      0    0
```

```
            features
docs           positive negative
  reagan1981          1        1
  nixon1973           1        1
```

Figure 2: From text to tokens to matrix.

equating a word with its *lemma*, or canonical (dictionary) form; and the elimination of words either through the use of predefined lists of words to be ignored or based on their relative infrequency. The first form of textual data pre-processing treats words as equivalent when they differ only in their inflected forms, so that, for example, the different words *taxes*, *tax*, *taxation*, *taxing*, *taxed*, and *taxable* are all converted to their word stem *tax*. The second common textual pre-processing practice is to remove words that are considered unlikely to contribute useful information for analysis. These words, commonly called *stopwords*, are usually function words such as conjunctions, prepositions, and articles that occur in the greatest frequency in natural language texts but add little specific political meaning to the text that would be deemed useful to analyse from textual data. (See Figure 2.) The problem with excluding words from an pre-set list, however, is that there exists no universally suitable list of words known to contribute nothing useful to any textual data analysis. For instance, the pronoun *her*, as Monroe, Quinn and Colaresi (2008) found, has a decidedly partisan orientation in debates on abortion in the U.S. Senate. For these reasons, it has been noted that a general trend in preparing textual data for analysis has been gradually to reduce or eliminate reliance on stop lists (Manning, Raghavan and Schütze, 2008, 27). Another approach to restricting the focus of textual data analysis from all words to only potentially informative words is to filter words by indices constructed from their relative frequency across as well as within documents, through a weighting or trimming scheme based on frequencies (discussed below), but this first requires a matrix of all eligible features to be formed.

Other methods of processing tokens include converting text to "*n*-grams," defined as sequences of *n* consecutive tokens to form not words but phrases. This is a brute force method of recovering politically meaningful *multi-word expressions* that might contain identical unigrams but as phrases, mean exact opposites, such as *economy* in the multi-word expressions *command economy* and *market economy*. Also known as *collocations*, such expressions can be detected by statistical methods (e.g. Dunning, 1993). Detecting specific multi-word expressions is generally preferable to simply forming all *n*-grams, since the *n*-gram approach increases the number of features by (nearly) a multiple of *n*, and most of these will occur very rarely or represent frequently occurring but uninteresting combinations such as *let us*.

Types represent unique words, but we should remember that this is typically based on their forming unique combinations of characters. Especially in English, *homographs* (words that are different but that spelled identically) will appear falsely as the same word type, at least to the machines we are using to process them. We could be more specific in distinguishing these by using a *part-of-speech* (POS) tagger that will at least distinguish homographs that are not the same parts of speech. In the example we cited earlier of the different uses of the term *kind*, for instance, a part of speech tagger could have annotated our tokens to distinguish these types (and this is indeed how I computed the proportions of its different forms in that example). Annotating tokens using a POS tagger can help us distinguish terms with opposite meanings such as *sanctions* in the sentence: *The President sanctions the sanctions against Iran*, by treating these as “sanctions/VERB” and “sanctions/NOUN”, one meaning permission and the other meaning a penalty. Despite the obvious advantages, however, differentiating word types using POS taggers in bag-of-words approaches to text as data is seldom, if ever, currently used.

It is not uncommon to read in a published application based on the analysis of text as data, perhaps in a footnote, that the authors took “the standard pre-processing steps” to prepare their input texts. In truth there is no standard, and without details of the specific steps a researcher took, such summary references as to what invasive procedures were applied to the text are uninformative. Each application will have different needs for feature processing, with different consequences as a result of the choices made at this stage. In one of the few systematic studies of feature processing choices and their consequences, Denny and Spirling (2018) replicated several published text analyses from political science using a variety of alternative feature processing steps. Their results shows that “under relatively small perturbations of of preprocessing decisions...very different substantive interpretations would emerge” (187). Researchers in practice should be aware of these decisions, critically examine the assumptions of their methods and how these relate to feature selection, and test the robustness of these results.

Converting of textual features into a quantitative matrix. This is mainly a mechanical step, resulting in a matrix whose dimensions are determined by the choices relating to the definitions of documents and features. We have already mentioned that some schemes call it a *document-term matrix*. (Some might even call it a *term-document matrix*, but there are great advantages in fixing the “documents” to be row units and saving our efforts to promote diversity for more important problems.) We have been using the term *feature* thus far, but it is worth noting why and how this is different from just speaking about “terms”. Computer scientists use *feature* to refer to what social scientists have long called *variables*: attributes of our units of analysis that differ across units. Because calling them *features* emphasises how they differ from terms or words (and may no longer even be words), I use this term to denote the selections and transformations made from token-based units that become the data used for analysis. I prefer the term *features* since the tokens have invariably been transformed in some way before they are shaped into a matrix, or may be abstractions from tokens such as annotations or dictionary keys rather than even transformed tokens.

Most matrices containing feature frequencies are characterised by a high degree of *sparsity*, meaning that they are mostly zeroes. Document-feature matrices are affected by what is known in machine learning as the *curse of dimensionality*: new observations also tend to grow the feature set, and each new term found in even a document adds a new column to the matrix. There is even a “law” named for this in linguistics: *Heap’s Law*, which states that the number of types grows exponentially with the number of tokens.[9] Forming a matrix of the (lower-cased) word features from the pre-2020 US presidential inaugural address corpus, for instance, creates a matrix of 58 inaugural speech documents by 9,273 features, but nearly 92 percent of the 537,834 cells in this matrix are zeros. In fact, over 41 percent of the features in this matrix are *hapax legomena*, defined as words that occur only a single time, such as the term *aborigines* in Ulysses S. Grant’s (politically incorrect) promise “to bring the aborigines of the country under the benign influences of education and civilisation.”

One strategy for mitigating the problem of exponentially increasing dimensionality is to trim or to weight the document-feature matrix. Trimming can be done on various criteria, but usually takes the form of a filter based on some form of feature frequency. Weighting schemes convert a matrix of counts into a matrix of weights. The most common of these is relative term frequency, a weighting process also known as document *normalisation* because it homogenises the sum of the counts for each document. Since documents in a typical corpus vary in length, this provides a method for comparing frequencies more directly than counts, which are inflated in longer documents (although these frequencies are also subject to length effects related to Heap’s law). Other popular weighting schemes are *tf-idf*, or term frequency-inverse document frequency, popular as a method in information retrieval for down-weighting the terms that are common to documents. In addition to the term frequency already discussed, *tf-idf* adds a weight that approaches zero as the number of documents in which a term appears (in any frequency) approaches the number of documents in the collection.[10] When we have selected our texts because they pertain to a specific topic — as we usually will — then inverse document frequency weighting means zeroing out most of our topical words, since these will appear in

[9] Technically speaking, Heap’s Law states that $M = kT^b$, where M is the vocabulary size (the number of unique word types), T is the number of tokens, and k and β are constants for computational linguists to estimate and argue about (but that are usually $30 \leq k \leq 100$ and $b \approx 0.5$). (Manning, Raghavan and Schütze, 2008, 88)

[10] Perhaps surprisingly, there is no universal definition of *tf-idf* weighting, and formulas may differ depending on whether the *tf* is a count or a proportion, what sort of constant may be added, or what logarithmic base and constant are applied to the inverse document frequency. A good measure, however, is $tf_{ij} * \log_{10} \frac{N}{df_j}$, where tf_{ij} is the count of feature j in document i, N is the number of documents in a collection, and df_j is the number of documents in which feature j occurs (Manning, Raghavan and Schütze, 2008, 118). A feature occurring in all N documents thus receives a weight of zero since $\log(1) = 0$.

most or all documents. In texts of debates over health care, for instance, *tf-idf* weighting is likely to eliminate all words related to health care, even when they might occur at very different rates across different documents. If we think that it is not the occurrence, but rather the relative frequencies of words that are informative, then using *tf-idf* weighting is the opposite of what we want. While it will automatically remove "stop words" without using a list, *tf-idf* weighting will also throw out the substantive baby with the linguistic bathwater. Except for classification tasks where removing all but the most discriminating features can improve performance, *tf-idf* weighting is usually inappropriate for the analysis of political texts.[11]

Because the rows and columns of the document-term matrix are unordered, the features that were originally carefully ordered words, in carefully ordered sentences, are now stored in a matrix object with no representation of order. In natural language processing, this approach is known as "bag of words", because it has disassociated the words from their context. For this reason, some text as data analyses use a different representation of documents based on token *vectors*, since these preserve order. For token vectors to be used in most analyses however, such as computing a similarity score between token counts, these need to be aligned into what is effectively a matrix representation. Other forms of analysis, such as forming co-occurrence matrixes, require iterating over the token streams and tabulating counts that are later combined into matrix form.

We have already noted the curious inter-relationship between features, and documents as collections of features. Some matrix representations do away with the notion of documents altogether, forming feature-by-feature matrixes counting how features co-occur within a defined context. This context might be the original document, or a moving local window for each target feature, for instance the five tokens found before or after the target feature. (Note here that I am very specifically using *token* to refer to a word when it exists as a segmented textual unit, but *feature* when it has been shaped into a matrix.) Known as a *feature co-occurrence matrix*, this matrix is a special variant of our document-feature matrix, where the documents have been redefined as features themselves, and the counts are tabulated within a context that we define. This is the basis for input into network analysis, for instance inter-relationships of words based on their co-occurrence.

For simplicity, the focus here is on features based on a bag-of-words approach, but matrix representations can be generalized to include weights based on word embedding vectors, possibly redefining documents as new units such as sentences or paragraphs. We have already mentioned the

[11] We could also add that many models commonly used in political science — such as the "Wordfish" Poisson scaling model or variants of Latent Dirichlet allocation (topic) models — only work with counts as inputs, so that *tf-idf* or other weighting schemes are inapplicable.

popularity of vector representations of term features estimated from word embedding models. One option at the stage of creating the document feature matrix is to combine the counts with weights or scores from these word vectors, especially with comparing documents (for semantic similarity, for instance) or for text classification using predictive models. Methods exist for combining word vectors with *tf-idf* weights to turn documents into more semantically meaningful matrix representations, extending the notion of the document-feature matrix into a more complex representation than the simpler version depicted in Figure 2.

Analysing the textual data using an appropriate quantitative or statistical procedure. The key here is *appropriate*: does the procedure for analysing the textual data produce reliable and valid insights into the question motivating the analysis? It is worth keeping in mind that by the time we have reached this stage of the analysis, we have already proceeded on the basis of some strong assumptions, namely:

1. The texts accurately represent the underlying target of concern.
2. Our sample of texts are a typical or at least complete representation of the phenomena that is our target of concern.
3. Our conversion of the texts into data has retained the essential information we need to provide insight on our target of concern.

The first assumption is by no means obvious in politics, where much verbal activity could be dismissed as “cheap talk” or as insincere promises or false or misleading claims,[12] but we have good reasons to think that text is more sincere than other forms of behaviour, especially in a legislative setting (Herzog and Benoit, 2015). Our selection from these also needs to be based on sound principles, just as data selection does in any research exercise. The third choice is something we have just discussed but involves many additional and deeper issues. It also interacts with a fourth strong assumption made at the analysis stage:

4. The analytic procedure yields a reliable and valid basis for inference on our target of concern.

The main risks with respect to reliability comes when human judgment forms part of either the process of extracting data features or performing the analysis. In content analysis, for instance, human coders may be responsible for both defining the units of textual data and for assigning them annotations (“codes”) based on their reading the textual units and judging the most applicable category from a set of instructions. The former process is known as *unitization* and the second as *coding* Krippendorff

[12] See Glenn Kessler, Salvador Rizzo, and Meg Kelly, “President Trump made 8,158 false or misleading claims in his first two years.” *Washington Post* January 21. `https://www.washingtonpost.com/politics/2019/01/21/president-trump-made-false-or-misleading-claims-his-first-two-years`

(2013) (although computer scientists typically call this text *annotation*). Both processes can pose severe challenges for even trained and highly educated human coders to apply at conventionally acceptable rates of reliability and inter-coder agreement Mikhaylov, Laver and Benoit (2012). With respect to the potential unreliability of the analytic procedure, this is seldom a problem in text as data designs, because even the simplest procedures — such as comparing the relative rates of negative versus positive sentiment words — are implemented computationally in ways that would not differ according to the judgment or personality of the analyst.

The validity of the analytic procedure in terms of providing insight on the target of concern is strongly influenced by the choices made at the feature extraction stage. Often, identical choices might be suitable for one analytic purpose but unsuitable for others. Consider the following three sentences, which we might wish to compare using a measure of textual similarity, such as *cosine similarity*, a measure that ranges (for text counts) between 0.0 to indicate the absence of any correlation and 1.0 to indicate two texts with identical feature proportions.

(a) *Party X embraces protection of citizens through universal health care.*
(b) *Party Y prioritises economic growth, even at the cost of environmental protection.*
(c) *Party Y prioritises environmental protection, even at the cost of economic growth.*

The cosine similarity of text a) with the second text is fairly low (at 0.32), as we might expect given that it concerns a different area of policy (but both are still statements about a party and both use the term *protection*). If we wanted to measure differences in the policy areas receiving *attention*, a measure of similarity based on vectors of word occurrences might suit our purposes well. But if we wished to differences in policy *position*, then cosine similarity in this example would be a poor instrument, as it indicates perfect similarity (1.0) between texts b) and c), despite these indicating exactly opposite political priorities. We can think of ways to differentiate them that might involve using sentence structure rather than simple bag-of-words approaches, but this only underscores the point that the appropriate choice of analytical procedure is influenced by choices made at the feature extraction stage.

Often there is an iterative process between the feature extraction and analysis stages, in which following a preliminary analysis, we need to return to the feature extraction and processing stage in order to make adjustments before repeating the analysis. Sometimes, this might result from examining unintended or anomalous results from an analysis and deciding that these would be better avoided through different feature processing choices. Observing clusters of the same root terms with different inflections, for instance, could motivate stemming the tokens and repeating the analysis. Likewise, anyone who has plotted a *word cloud* of unselected features (where these appear in sizes proportional to their relative frequency) will quickly return to more aggressive feature selection choices when they see the words *the* and *and* dominating the plot. Other feature processing decisions can especially influence unsupervised methods such as topic models, because unsupervised approaches necessarily attempt to learn from all supplied information. Or, observing a set of topics sharing high proportions of stopwords might be cleaned up by removing stopwords from features prior to fitting the topic model (and removing stopwords almost always improves the interpretability of topics fit using topic models). In their study of the effects of these choices, Denny and Spirling (2018, 187) found that key "modelling choices, such as the optimal number of topics, were also startlingly dependent" on decisions made at the feature processing stage. Other techniques may be more robust to this, especially supervised methods or those that automatically down-weight uninformative features through their conditional probabilities or by applying a regularization penalty. The best fine-tuning will be a combination of theoretically motivated choices of feature processing, confirmed by careful inspection following analysis.

Should we be concerned that this cycle might encourage dishonesty, by tweaking our feature extraction until we get the results we want? In short, no, although of course we should not contrive results. Residual diagnostics have long been a feature of basic statistical analysis, and often these serve to detect anomalies that indicate errors to be corrected before re-running the results, or fixes to be applied to get our data to conform more closely to the assumptions of our model (such as applying a log transformation to skewed variables or applying weights to heteroskedastic residuals in least-squares regressions). In working with textual data, this process is all the more important. Natural language often shows a slippery resistance to neat transformation into data, because of features such as polysemy or the fact that many words in non-compounding languages lose an important part of their meaning when separated from the multi-word expressions in which they occur. Or, it might be a simple matter of spelling or OCR mistakes indicating we have a cluster of words that should be the same but whose characters need correcting because an "i" was rendered as an "l" or a "o" as a "0", or because we did not remove running page footer from texts converted from pdf format. We should never underestimate just how messy can be the process of converting text, no matter how clearly we can read it, into clean features of textual data. Often, the best — or the first — stage at which this becomes fully apparent is during analysis, and when detected it often means returning to earlier stages, cleaning things up or making better choices, and repeating the analysis. This is a far more valid and honest approach than sticking with results that we know are wrong, and could have fixed, had we only gotten cleaner electronic texts to begin with or had been better informed about the full consequences of our feature processing decisions.

Many of the analytic procedures we apply to textual features take the form of advanced statistical models that impose strong assumptions on the data generating process, such as assuming that conditional word counts are identical and independently distributed as a Poisson (e.g. Slapin and Proksch, 2008), a negative binomial (e.g. Lo, Proksch and Slapin, 2016), or a multinomial process (e.g. Roberts et al., 2013). We know with certainty that words are not conditionally or positionally independent and that the degree of non-compliance will vary from mild to extreme in non-systematic ways, depending on the stylistic choices of a speaker or writer as well as characteristics of the language being used. To apply the tested and well known properties of statistical data analysis to text, we must impose assumptions about the data-generating and stochastic processes that come with statistical approaches. The problem is, there exists no neat, parsimonious model of the data-generating process for natural language, so we rely on models whose assumptions are violated in sometimes painfully obvious ways. Fitting models that violate statistical assumptions is hardly new in social science, but because we so directly and intimately understand the nature of the source data (natural language) we are likely to be more acutely aware of these problems.

The good news is that even when violating statistical assumptions wholesale, we still get a tremendous amount of useful juice from models that are highly simplistic from a linguistic point of view. The "naive" in "Naive Bayes", after all, is an overt recognition that its class conditional probabilities are wrong, because the conditional independence required to compute the joint probabilities from word counts is blatantly fictional. Yet, Naive Bayes remains a highly useful tool for classifying texts (Zhang, 2014). It is hard to summarise this better than have Grimmer and Stewart (2013, 4):

> The complexity of language implies that all methods *necessarily* fail to provide an accurate account of the data-generating process used to produce texts. Automated content analysis methods use insightful, but wrong, models of political text to help researchers make inferences from their data.. . . Including more realistic features into quantitative models does not necessarily translate into an improved method, and reducing the assumptions used may not imply more productive analyses. Rather, subtleties of applying the methods to any one data set mean that models that are less sophisticated in the use of language may provide more useful analysis of texts.

Two additional considerations often guide our choice of analytical method for analysing the features of textual data. One is interpretability, something we discuss more in our final stage. A second consideration is computational efficiency. Even with cheap, efficient computing resources, some models can be enormously expensive to fit. The advantages in low computational cost of fitting simpler, efficient models such as linear SVMs or Naive Bayes might well outweigh marginal gains in classifier performance from more advanced, but more computationally expensive models such as recurrent or convolutional neural network models. In addition, simpler methods often prove more robust in the sense of avoiding overfitting, a risk which every computer scientist acknowledges but which few explore in published applications (which typically aim to demonstrating how a new method has outperformed all other methods at some specific task for a specific dataset). As social scientists, we must give far greater priority to robustness and its transparent demonstration in our choice of method for analysing text as data.

Interpreting and summarising the results. Summarising and communicating findings is the end stage of any analysis, and the analysis of text as data is no different. Because it involves making sense following abstraction and analysis of raw input data that we could make direct sense of to begin with, however, interpreting the results of textual data analysis can involve some special challenges. Because the analytical stage involved using a trusted methodology, we typically stake our claim of validity of results on the basis that they inherit the trusted procedural properties of the methodology. But because the application of text analysis methods always involves choices at earlier stages, there is an additional measure of trust to establish upon interpreting results, namely that the researcher has appropriately processed the texts and correctly applied the analytic method. This is usually established through additional results showing the robustness of the conclusions to different choices or demonstrating that the parameters of one's model (such as the number of topics) are optimal. Robustness checks are common in econometric analyses of non-textual data, but only recently have begun to form parts of textual data analyses in the social sciences.

Especially when text analysis is exploratory, such as demonstrating a new application or methodology, validation is a crucial part of interpreting one's results. For supervised scaling methods, this is tricky because there is seldom an objective measure to which text-based point estimates can be compared. Instead, we typically rely on comparison to some external measures obtained through alternative, often non-textual means, such as expert survey estimates of policy positions in the case of scaling ideology. Validating supervised classification methods is easier, because we could have objective labels for verifying predicted classes (such as observed party affiliation), numeric scores (from a survey question), or labels assigned through human annotation.

Interpreting the results of unsupervised methods is trickier, because it often involve reading into the textual contents of topics or word weights and deciding whether they accord with some reasonable interpretation of the world. Point estimates from unsupervised scaling can be compared to the same sorts of external measures as supervised scaling estimates, or to summaries of detailed human readings of the scaled texts (e.g. Lowe and Benoit, 2013). Topic models are trickier, but typically involve reading the word features that are the most frequent in each topic and assigning a label to that topic. Roberts et al. (2014, 1073) for instance interpreted their "Topic 1" as "the 'crime' and 'welfare' or 'fear' topic", because its most frequently used (stemmed) word features included *illeg*, *job*, *immigr*, *welfar*, and *crime*. Their second topic, which they interpreted as emphasising "the human elements of immigrants", also contained among its most frequent words *immigr*, *illeg*, *border*, and *worri*. These distinctions are hardly clear-cut, and any labels attached to topics is ultimately subjective. Model-based diagnostic for setting an optimal number of topics, furthermore, may be unrelated or even negatively correlated with topics' semantic coherence (Chang et al., 2009). The best application of unsupervised methods will produce results that are semantically coherent with our understandings of the texts and with the world that our analysis of them aims to represent.

As analytical tools become increasingly sophisticated, we now have access to powerful methodologies whose procedural workings may be non-transparent. No one has figured out the data-generating process of language, but with modern approaches for classification, this has become unnecessary. Some of the best performing classification methods for text, for instance, use "deep learning" models fit to the level of characters. When fed with enormous amounts of data, convolutional neural network models can outperform other approaches (Zhang, Zhao and LeCun, 2015) but it impossible to assess their operation in any application in the way one would diagnose even an advanced computational method such as fitting a Markov chain Monte Carlo model. Because social science eschews black boxes, we often stick to interpretable models even when better-performing alternatives exist, especially if we have large quantities of data. If a huge amount of training data are available, "then the choice of classifier probably has little effect on your results" (Manning, Raghavan and Schütze, 2008, 337), and we should be guided by the principle of parsimony to prefer more transparent and simpler models over more opaque and complex ones, even at the cost of small tradeoffs in performance. Just as our concern in social science is explanation rather than prediction, we generally prefer model specification based on theory and isolating the effect of specific explanatory factors, rather than attempting to include every possible variable to maximise variance explained. Because the goals of explanation or measurement differ from the (typical machine learning) objective of prediction, it is worth reminding ourselves of this preference.

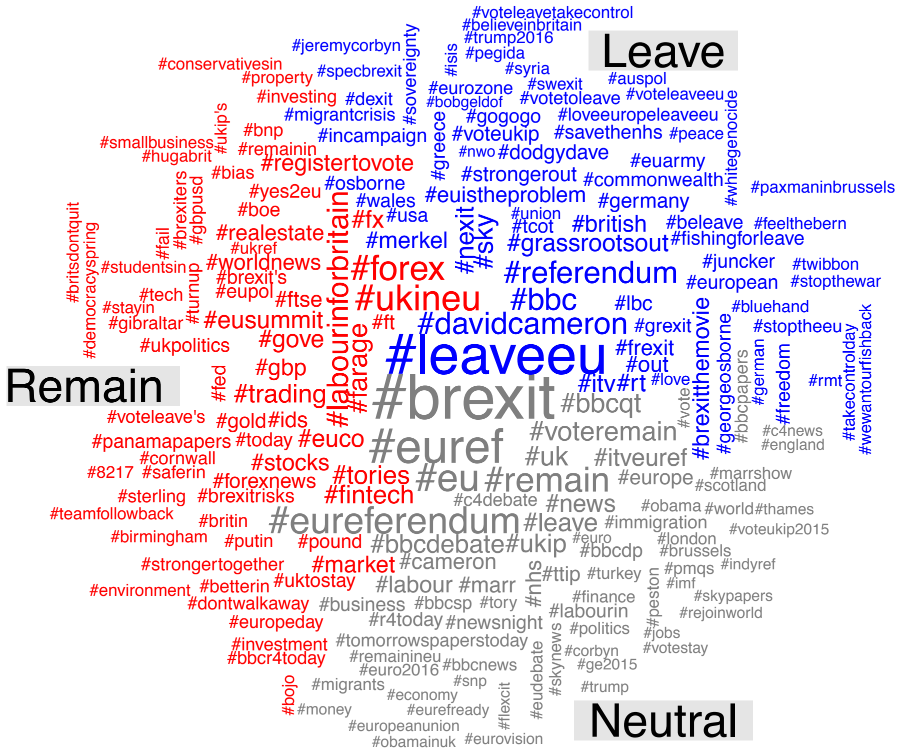


Figure 3: Word cloud of influential hashtags from a sample of Tweets about Brexit. From Amador Diaz Lopez et al. (2017).

Communicating the results of text analysis in a compact and effective way is practically challenging because numerical tables only poorly capture the full nuances of language, and we typically have too many features and documents (or topics) to fit these easily into a format that will not overwhelm a reader. Graphical presentation of text analytic results is especially important, and should offer special opportunities given that we can read and interpret word features when they form the elements of a plot. Despite this potential, however, innovation in visual presentation of text analytic results has been slow to non-existent, moving little beyond the "word cloud" and its variations. Designed to show the most frequent terms, the word cloud plots word features in sizes proportional to their relative frequency in the textual dataset, producing a plot with some visual appeal but often no clear communication of any particular result. This is slightly improved by using a "comparison" word cloud that partitions word plots according to external categories, such as the Twitter hashtags used in Figure 3 according to whether the user was predicted to support Brexit or not (Amador Diaz Lopez et al., 2017). Other methods exist of course, especially for characterizing the semantic content of topics from topic models, probably the area where the most innovations of visual presentation in text analysis have occurred (e.g. Figure 5 from Reich et al., 2015). Given the unique interpretability of word features, however, it is justified to feel that we should have developed more imaginative graphical ways to include words in our plots (and not just on the axis labels).

A final word on presentation and interpretation concerns how we characterize the *uncertainty* of our text analytic results. In addition to inheriting procedural validity established by decades of statistical theory, the quantitative analysis of text as data also makes it possible to quantify the uncertainty of our results. In the analysis of text as data, this can take two forms, parametric and non-parametric. Parametric methods rely on the assumptions we have imposed on the data through some model of its stochastic process, in the context of an established procedure for producing estimates — such as maximum likelihood or simulations from Bayesian posterior distributions. Uncertainty estimates produced by these methods are typically too small because of unmodeled heterogenieity in our model of text data, but even this bias can be quantified. Another approach is non-parametric, through through bootstrapping a text by resampling from its elements. In exploring different methods of charactering uncertainty for measurement models of text, Lowe and Benoit (2013) advocate repeating the analysis with different versions of a text that have been reassembled after resampling their sentences with replacement, to create hypothetical variations of documents drawn from the sentences of the documents actually observed. This method has begun to appear in different applications, such as Benoit, Munger and Spirling (2019) who use it to compute confidence intervals for document-level readability statistics, but it has been slow to catch on despite its almost universal applicability to text as data analyses. Measuring uncertainty in the analysis of text as data remains one of the most important challenges in this field (Grimmer and Stewart, 2013, 28), and should be a requirement if we are to give the quantitative analysis of text full methodological status alongside that of non-textual data.

5 Conclusions and Future Directions

Treating text as data means converting it into features of data and analysing or mining these features for patterns, rather than making sense of a text directly. This process turns text from something directly meaningful into data that we cannot interpret in its raw form, but whose analysis produces meaningful insights using structured rules in ways and at scales that would otherwise be impossible. This approach to text analysis has become increasingly mainstream in the social sciences, and the methods and applications increasingly innovative. This trend, which is likely to continue, has been driven by several forces.

First, as in so many other areas of human activity, in textual analysis the rise of the machines has enabled scholars to automate key parts of the analytic process, a process formerly performed using qualitative methods by unreliable humans who actually knew what they were doing. With text analytic methods, humans can now mine large quantities of textual data, using sophisticated methods, implemented by perfectly reliable computer automatons.

A second force driving the textual revolution has been one of scale: the incredible volume of texts available today requires automated, quantitative approaches if we are to analyse more than a small subset of this data. The growth of electronic publications has made machine-readable text ubiquitous, and along with it the promise of a huge wealth of information about the characteristics of the political and social actors that generate these texts. Such texts include legislative speeches, political party manifestos, legal decisions, election campaign materials, press releases, social arguments, correspondence, and television and radio transcripts, to name but the key ones. Resource limitations may still cause us to sample from available texts, but this involves much larger samples than in previous eras. Miners want to extract all, not just a sample, of the rich resources available, and the logic of text mining points to the same conclusion. Methods that require reading a text, or determining what it "means", are simply not applicable to a scale of tens or hundreds of thousands or more texts. Instead, we need tools that can turn unstructured text into structured information, using inexpensive and efficient methods for parsing, annotating, and categorising the elements of text to prepare it for analysis and then to perform this analysis.

Of course, a need to process big textual data and the machines to implement this processing are only as useful as the methodologies that the machines can implement. A final enabler (and driver) of the shift to treating text as data has been the development and application of sophisticated statistical learning methods for extracting information and generating inferences from textual data. These are extensions of statistics and machine learning but with specific applications to textual data.

Many challenges lie on the road ahead, and these should be met in the same way as most other breakthroughs in social science methodology: through innovations required to solve specific research problems as part of our agenda to understand the political and social world. Some challenges have already been identified, such as a need for more validation of our models of textual data under a broader range of circumstances, and a more realistic way to characterise uncertainty in textual data analysis. Some are just emerging, such as how to incorporate named entity recognition and part-of-speech tagging to distinguish alternative meanings, or how to identify and make use of multi-word, non-compositional phrases (how to distinguish, in other words, *Homeland Security* from *social security*). Other recent innovations include the merger of human qualitative input for processing textual data with statistical scaling or machine learning for analysis, possibly using crowd-sourcing for text annotation through an agile, “active learning” process. As advances continue in other fields such as machine learning and natural language processing in computer science, we must keep a firm grip on political science research objectives and standards while at the same time borrowing what is useful to our discipline. As we gain experience and understanding through both theory and applications, textual data analysis will continue to mature and continue to produce valuable insights for our understanding of politics.

References

A, Hertel-Fernandez and K. Kashin. 2015. "Capturing business power across the states with text reuse." Presented at the Annual Meetings Midwest Political Science Association, Chicago, April 16–19.

Amador Diaz Lopez, Julio Cesar, Sofia Collignon-Delmar, Kenneth Benoit and Akitaka Matsuo. 2017. "Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data." *Statistics, Politics and Policy* 8(1):210–20.

Baek, Young Min, Joseph N Cappella and Alyssa Bindman. 2011. "Automating Content Analysis of Open-Ended Responses: Wordscores and Affective Intonation." *Communication Methods and Measures* 5(4):275–296.

Baumgartner, F. R., C. Green-Pedersen and B. D. Jones. 2008. *Comparative Studies of Policy Agendas*. Routledge.

Benoit, Kenneth, Kevin Munger and Arthur Spirling. 2019. "Measuring and Explaining Political Sophistication Through Textual Complexity." *American Journal of Political Science* 63(2):491–508.

Benoit, Kenneth, Michael Laver, Christine Arnold, Madeleine O. Hosli and Paul Pennings. 2005. "Measuring National Delegate Positions at the Convention on the Future of Europe Using Computerized Wordscoring." *European Union Politics* 6(3):291–313.

Berelson, Bernard. 1952. *Content analysis in communications research*. New York: Free Press.

Blei, D.M., A.Y. Ng and M.I. Jordan. 2003. "Latent dirichlet allocation." *The Journal of Machine Learning Research* 3:993–1022.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama and Adam T Kalai. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." pp. 4349–4357.

Budge, Ian, David Robertson and Derek Hearl. 1987. *Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press.

Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and Eric Tanenbaum. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*. Oxford: Oxford University Press.

Caliskan, Aylin, Joanna J. Bryson and Arvind Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356(6334):183–186.
URL: *https://science.sciencemag.org/content/356/6334/183*

Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang and D. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*.

Chilton, Paul. 2017. "'The people' in populist discourse: Using neuro-cognitive linguistics to understand political meanings." *Journal of Language and Politics* 16(4):582–594.

Colleoni, Elanor, Alessandro Rozza and Adam Arvidsson. 2014. "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data." *Journal of Communication* 64(2):317–332.

Covington, M.A. and J.D. McFall. 2010. "Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR)." *Journal of Quantitative Linguistics* 17(2):94–100.

Däubler, Thomas and Kenneth Benoit. 2018. “Estimating Better Left-Right Positions Through Statistical Scaling of Manual Content Analysis.” London School of Economics typescript.

Däubler, Thomas, Kenneth Benoit, Slava Mikhaylov and Michael Laver. 2012. “Natural Sentences as Valid Units for Coded Political Texts.” *British Journal of Political Science* 42(4):937–951.

Denny, Matthew J and Arthur Spirling. 2018. “Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It.” *Political Analysis* 26(2):168–189.

Dunning, Ted. 1993. “Accurate methods for the statistics of surprise and coincidence.” *Computational linguistics* 19:61–74.

Fairclough, Norman. 1992. “Intertextuality in critical discourse analysis.” *Linguistics and Education* 4(3-4):269–293.

Fairclough, Norman. 2001. *Language and power*. Pearson Education.

Fang, Xing and Justin Zhan. 2015. “Sentiment analysis using product review data.” *Journal of Big Data* 2(5):1–14.

Firth, John R. 1957. A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*. Oxford: Blackwell pp. 1–32.

Flesch, Rudolph. 1948. “A new readability yardstick.” *Journal of Applied Psychology* 32(3):221–233.

Foucault, Michel. 1972. *The Archaeology of Knowledge and the Discourse on Language*. New York: Pantheon Books.

Greenacre, Michael. 2017. *Correspondence analysis in practice*. Chapman and Hall/CRC.

Grimmer, J. 2010. “A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases.” *Political Analysis* 18(1):1–35.

Grimmer, 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):267–297.

Gunning, Robert. 1952. *The Technique of Clear Writing*. New York: McGraw-Hill.

Herdan, Gustav. 1955. “A new derivation and interpretation of Yule’s ‘Characteristic’ K.” *Zeitschrift für angewandte Mathematik und Physik ZAMP* 6(4):332–334.

Herzog, Alexander and Kenneth Benoit. 2015. “The Most Unkindest Cuts: Speaker Selection and Expressed Government Dissent during Economic Crisis.” *The Journal of Politics* 77(4):1157–1175.

Jagers, Jan and Stefaan Walgrave. 2007. “Populism as political communication style: An empirical study of political parties’ discourse in Belgium.” *European Journal of Political Research* 46(3):319–345.

Joachims, Thorsten. 1999. Transductive inference for text classification using support vector machines. In *ICML*. pp. 200–209.

Klemmensen, Robert, Sara Binzer Hobolt and Martin Ejnar Hansen. 2007. “Estimating policy positions using political texts: An evaluation of the Wordscores approach.” *Electoral Studies* 26(4):746–755.

Klüver, Heike. 2009. “Measuring Interest Group Influence Using Quantitative Text Analysis.” *European Union Politics* 10(4):535–549.

Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology*. 3rd ed. Thousand Oaks, CA: Sage.

Labbé, Cyril, Dominique Labbé and Pierre Hubert. 2004. "Automatic Segmentation of Texts and Corpora." *Journal of Quantitative Linguistics* 11(3):193–213.

Lai, Siwei, Liheng Xu, Kang Liu and Jun Zhao. 2015. Recurrent Convolutional Neural Networks for Text Classification. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Landauer, Thomas K, Peter W Foltz and Darrell Laham. 1998. "An introduction to latent semantic analysis." *Discourse processes* 25(2-3):259–284.

Lasswell, Harold Dwight. 1948. *Power and Personality*. New York: W. W. Norton and Company.

Laver, Michael and Kenneth Benoit. 2002. "Locating TDs in Policy Spaces: Wordscoring Dáil Speeches." *Irish Political Studies* 17(1):59–73.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Estimating the policy positions of political actors using words as data." *American Political Science Review* 97(2):311–331.

Leites, Nathan, Elsa Bernaut and Raymond L Garthoff. 1951. "Politburo images of Stalin." *World Politics* 3(3):317–339.

Liu, Dilin and Lei Lei. 2018. "The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election." *Discourse, context & media* 25:143–152.

Lo, James, Sven-Oliver Proksch and Jonathan B. Slapin. 2016. "Ideological Clarity in Multiparty Competition: A New Measure and Test Using Election Manifestos." *British Journal of Political Science* 46(3):591–610.

Loughran, Tim and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66(1):35–65.

Lowe, William and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits From Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21(3):298–313.

Lucas, C, R A Nielsen, M E Roberts, B M Stewart, A Storer and D Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* pp. 1–24.

Manning, C. D., P. Raghavan and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Martindale, Colin. 1975. *Romantic progression: The psychology of literary history*. Hemisphere publishing corporation.

Mikhaylov, Slava, Michael Laver and Kenneth Benoit. 2012. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20(1):78–91.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pp. 3111–3119.

Monroe, B. and K. Maeda. 2004. "Talk's Cheap: Text-Based Estimation of Rhetorical Ideal-Points." POLMETH Working Paper.

Monroe, B. L., K. M. Quinn and M. P. Colaresi. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16(4):372–403.

Monroe, Burt L and Philip A Schrodt. 2009. "Introduction to the Special Issue: The Statistical Analysis of Political Text." *Political Analysis* 16(04):351–355.

Pang, B., L. Lee and S. Vaithyanathan. 2002. “Thumbs up? Sentiment classification using machine learning techniques.” *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 79–86.

Pennington, J, R Socher and C D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Empiricial . . . .*

Perry, Patrick O and Kenneth Benoit. 2018. “Scaling Text with the Class Affinity Model.” *arXiv preprint arXiv:1710.08963* .

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer. 2018. “Deep contextualized word representations.”.

Proksch, S.-O. and J. B. Slapin. 2010. “Position taking in the European Parliament speeches.” *British Journal of Political Science* 40(3):587–611. forthcoming.

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

Reich, Justin, Dustin H Tingley, Jetson Leder-Luis, Margaret Roberts and Brandon Stewart. 2015. “Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses.” *Journal of Learning Analytics* 2(1):156–184.

Reisigl, Martin. 2008. Analyzing political rhetoric. In *Qualitative Discourse Analysis in the Social Sciences*, ed. Ruth Wodak and Michał Krzyżanowski. Basingstoke: Palgrave Macmillan pp. 96–120.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. “Structural Topic Models for Open-Ended Survey Responses.” *American Journal of Political Science* 58(4):1064–1082.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi et al. 2013. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. pp. 1–20.

Sanders, James, Giulio Lisi and Cheryl Schonhardt-Bailey. 2017. “Themes and Topics in Parliamentary Oversight Hearings: A New Direction in Textual Data Analysis.” *Statistics, Politics and Policy* 8(2):153–194.

Schonhardt-Bailey, Cheryl. 2003. “Ideology, party and interests in the British Parliament of 1841–47.” *British Journal of Political Science* 33(4):581–605.

Schonhardt-Bailey, Cheryl. 2008. “The Congressional Debate on Partial-Birth Abortion: Constitutional Gravitas and Moral Passion.” *British Journal of Political Science* 38:383–410.

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. “A Scaling Model for Estimating Time-Series Party Positions from Texts.” *American Journal of Political Science* 52(3):705–722.

Slapin, Jonathan B. and Sven-Oliver Proksch. 2014. Words as data: Content analysis in legislative studies. In *The Oxford Handbook of Legislative Studies*. Oxford: Oxford University Press pp. 126–144.

Spache, George. 1953. “A New Readability Formula for Primary-Grade Reading Materials.” *The Elementary School Journal* 53(7):410–413.

Spirling, Arthur. 2016. “Democratization of Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915.” *Journal of Politics* 78(1):120–136.

Steyvers, Mark and Tom Griffiths. 2007. Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*, ed. Thomas K. Landauer, Danielle S. McNamara, Simon Dennis and Walter Kintsch. Vol. 427 New York: Routledge pp. 424–440.

Stone, Philip J, Dexter C Dunphy and Marshall S Smith. 1966. *The General Inquirer: A computer approach to content analysis.* Cambridge, MA: MIT press.

Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis." *Journal of the American Statistical Association* 108(503):755–770.

Tausczik, Y R and James W Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29(1):24–54.

van Dijk, Teun. 1997. "What is political discourse analysis." *Belgian Journal of Linguistics* 11(1):11–52.

van Dijk, Teun A. 1994. "Critical discourse analysis." *Discourse & Society* 5(4):435–436.

Walgrave, Stefaan and Knut De Swert. 2007. "Where does issue ownership come from? From the party or from the media? Issue-party identifications in Belgium, 1991-2005." *The Harvard International Journal of Press/Politics* 12(1):37–67.

Wilkerson, J, D Smith and N Stramp. 2015. "Tracing the flow of policy ideas in legislatures: A text reuse approach." *American Journal of Political Science* 59(4):943–956.

Wilkerson, John and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20:529–544.

Wimsatt, William K and Monroe C Beardsley. 1946. "The intentional fallacy." *The Sewanee Review* 54(3):468–488.

Yu, Bei, Stefan Kaufmann and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech." *Journal of Information Technology & Politics* 5(1):33–48.

Zhang, H. 2014. The optimality of naive Bayes. Vol. 1 American Association for Artificial Intelligence.

Zhang, Xiang, Junbo Zhao and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in neural information processing systems*. pp. 649–657.