Computer assisted text analysis for comparative politics*

Christopher Lucas, Richard Nielsen, Margaret E. Roberts,
Brandon M. Stewart, Alex Storer, Dustin Tingley†

This draft: November 23, 2014

*Our thanks to Sam Brotherton and Jetson Leder-Luis for research assistance and Amy Catilinac for discussion about text analyses in comparative politics. We also thank Christopher Blattman, Dan Corstange, Macartan Humphreys, Amaney Jamal, Gary King, Helen Milner, Tamar Mitts, Brendan O'Connor, Arthur Spirling, and the Columbia University Comparative Politics Workshop for comments.

†Send comments to corresponding author: dtingley@gov.harvard.edu. Our software discussed in this paper is open source and available.

Abstract

Recent advances in research tools for the systematic analysis of textual data are enabling exciting new research throughout the social sciences. For comparative politics scholars who are often interested in non-English and possibly multilingual textual datasets, these advances may be difficult to access. This paper discusses practical issues that arise in the the processing, management, translation and analysis of textual data with a particular focus on how procedures differ across languages. These procedures are combined in two applied examples of automated text analysis using the recently introduced Structural Topic Model. We also show how the model can be used to analyze data that has been translated into a single language via machine translation tools. All the methods we describe here are implemented in open-source software packages available from the authors.

1 Introduction

In this paper we focus on new tools for comparativists to utilize *textual* data that can come in many different languages. Massive amounts of textual data are now available to comparativists, from debates in legislative bodies to newspapers to online social media. But using automated content analysis for comparative politics presents important challenges and opportunities, including processing and analyzing text in multiple languages and incorporating data we have about our texts directly into our analyses.

Comparativists are not unfamiliar with tools for textual analysis. Many of the automated text analysis innovations within political science were developed by comparativists (e.g. Schrodt and Gerner, 1994; Laver et al., 2003; Slapin and Proksch, 2008). After briefly orienting the reader to the range of text analysis methods available, we highlight a particular approach, unsupervised topic modeling. For interested readers, an online appendix provides an extensive discussion of supervised, scaling, and unsupervised methods to help readers understand the differences between existing approaches and identify methods that will be helpful for their own projects.

To showcase the potential of topic modeling for comparative politics we use the Structural Topic Model (STM) (Roberts et al., 2014, 2013, nd)to analyze Arabic fatwas and a novel multilanguage analysis of social media responses in Arabic and Chinese to the Edward Snowden event in June 2013. We argue in this paper that the STM should be an important part of the text analysis tool kit for comparativists. The STM provides a flexible way to incorporate "metadata" associated with the text, such as when the text was written, where (e.g., which country) it was written, who wrote it, and characteristics of the author, into the analysis using document-level covariates. In turn, it allows comparativists to understand relationships between metadata and topics in their text corpus.

A additional contribution of this paper is to discuss a range of tools that are necessary to analyze text from different languages. This includes a discussion of how text processing can differ across languages, along with discussion of robust software tools that properly account for differences across languages. We also consider how to simultaneously analyze

text in different languages. In doing so we discuss multilingual approaches to text analysis, briefly introduce a new `R` package, `translateR`, to access the Google and Microsoft machine translation APIs, and present a novel way to use the Structural Topic Model in a multilingual setting.

The structure of the paper is as follows. Section 2 discusses *research questions* in comparative politics that have benefited from text analysis tools, a multi-language view of *text processing*, and new tools for machine translation. Section 3 presents a brief review of *text analysis tools* with a particular focus on multi-language text modeling, and introduces the basics of the Structural Topic Model. Section 4 provides two example analyses using the Structural Topic Model. The first looks at Islamic fatwas and the second illustrates a novel way to use the model on machine translated data, with and application to social media responses in Arabic and Chinese to the Edward Snowden event.

2 Text and Language Basics

2.1 Research Questions and Data Analysis

Automated content analysis and comparative politics are well suited for each other. Countries around the world are producing textual data at unprecedented rates. Traditional government statistics are often missing, mis-measured, or manipulated, creating a strong incentive for scholars to turn to other forms of data. Meanwhile governments in almost all countries produce and store large amounts of text data that can be used for descriptive and causal inference. As internet connectivity rises, documents produced by individual citizens are becoming available from an increasingly diverse set of countries. E-mail and advances in survey technologies allow researchers to more easily collect interviews from politicians and government officials, expanding researchers' collections of qualitative data. The digitization of archives, historical records and public documents have exposed the inner workings of governments across the globe to the public eye.

While other disciplines are only recently catching on to text as a data source, scholars in comparative politics have been using text as data for years, and have built up intuitions for how text should be used for scholarly inference. Scholars of comparative politics have

drawn information from archives and interviews and therefore know how to ask political questions with this data, select important text or interview questions, and find meaningful patterns within the data (George and Bennett, 2005; Brady and Collier, 2010).

Scholars in comparative politics have already begun using automated methods for analyzing text to ask important political questions. Perhaps the most readily available form of text on politicians, scholars have been using records of speeches politicians make or deliberations among politicians to better understand the internal political workings of governments. Stewart and Zhukov (2009) use public statements by Russian leaders to understand how military versus political elites influence Russia's decision to intervene in neighboring countries. Baturo and Mikhaylov (2013) use federal and sub-national legislative addresses in Russia to identify leadership patterns within the Russian government. Schonhardt-Bailey (2006) uses a text clustering method to analyze thousands of pages of parliamentary debates in England to analyze the discussion about the repeal of the Corn Laws in Britain. Eggers and Spirling (2011) use parliamentary debates to model exchanges among politicians in the British House of Commons. Miller (2013) analyzes speeches in the United Nations to show that speeches by delegations from countries that were previously colonized devote more words to themes of victimization than states that were never colonized.

Others have tried to infer the policy positions of political parties or political leaders based on documents describing their positions on policies. The Comparative Manifestos project has collected electoral manifestos from all over the world, allowing scholars to use this text data to answer comparative questions about political systems (Budge, 2001). Early versions used human coding, but more recently the Comparative Manifestos project and related projects have been assisted by computer techniques. Catalinac (2013) uses thousands of Japanese election manifestos from 1986 to 2009 to determine how electoral strategies shifted after Japan's electoral reform in 1994. Nielsen (2012) uses *fatwas* from websites of Muslim clerics to measure the level of Jihadist thought in these clerics' writings and understand the drivers of Jihadism.

Political scientists have studied newspapers in various languages to ask questions

about media freedom and infer relationships between politicians and groups within a country. Van Atteveldt et al. (2008) analyze Dutch newspapers and extract relationships among political leaders and groups. Coscia and Rios (2012) use news to measure criminal activity in Mexico. Stockmann (2012) studies Chinese newspapers to study how media marketization influences anti-American sentiment in the Chinese media.

Finally, scholars in comparative politics have used blogs and social media sources. King et al. (2013) study the focus of censorship in social media in China, Jamal et al. (nd) study anti-Americanism in Arabic-language Twitter posts, and Barberá (2012) uses Twitter posts to scale citizen liberal-conservative ideal points across the US and several European countries. These papers demonstrate an emerging trove of data, being generated around the world. With more and more political discourse happening in these forums, comparative politics will require tools that can handle large volumes of data and systematic frameworks to analyze the data.

2.2 Text Processing Basics: A Multi-language View

In order to use automated methods to analyze text, first the analyst must ensure the text is machine-readable. Statistical methods for text analysis are often language agnostic, but the tools for pre-processing the texts are not. This can be challenging for newcomers in comparative politics as introduction to text analysis often focus exclusively on methods and software for English texts. We discuss three challenges that must be overcome that are particularly important when working with multiple languages within or across research project: dealing with encodings, pre-processing for dimensionality reduction, and handling of large corpora. Along the way we point out language-specific variations that comparativist studying particular countries should consider. In order to focus our discussion on less well known issues that come up when working outside of English, we leave to the Online Appendix A a more general discussion of topics that are more basic, such as the use of Optical Character Recognition. We discuss how we follow these procedures within our sections where we give examples.

2.2.1 Dealing with Encodings

The encoding of text is the way in which the computer translates individual, unique characters into bytes. Each language can have multiple encodings[1] and different computers and different software will default to recognizing different encodings. If the analyst is pulling data from multiple different sources, such as different webpages, it is likely that the text will be in different encodings. In this case, it is necessary to convert each document so that all of the encodings match.[2] The second step is to make sure the software reads the encoding correctly. This can often be done by changing the preference of the software, or encoding the text so that it matches the software's default encoding.

2.2.2 Pre-processing to Extract the Most Information

Automated text analysis methods usually treat documents as a vector containing the count of each word type within the document, disregarding the order in which the words appear. This 'bag-of-words' assumption reduces the dimension of natural language text, representing each document as a single vector with length equal to the number of unique words in the text. Unfortunately, even these dictionaries can be too large to be practical, ranging from thousands to millions of unique words. Fortunately, because most words appear only a few times in the corpus, removing infrequently occurring words can dramatically reduce the number of unique word types while having only a small impact on the number of tokens. Bounding the size of the vocabulary can play an important role in helping methods to perform well in practice.

In this section, we describe the most common tools for pre-processing textual data including stop word removal, stemming, lemmatization, compounding, decompounding, and segmentation. In each case the goal is to reduce the scale of the problem by treating words with very similar properties identically and removing words that are unnecessary to our interpretation and our model. Along with disregarding word order, the so called

[1]For example Chinese has several dozen encodings, the largest of which are Guobiao (GB), which has a two or four byte encoding, Big5 which has a one or two byte encoding, and ISO-2022, which has a seven byte encoding.

[2]Most programming languages have packages to transfer between encodings. For example, to convert encodings we use Python's package *chardet*.

'bag of words' assumption,[3] these procedures are common pre-processing steps but can differ across languages.

Stop word removal To aid in interpretation and model performance, analysts often remove words that are extremely common but unrelated to the quantity of interest. These "stop words" are dropped before the analysis. In most settings this involves removing frequently occurring function words such as "and" and "the", but often removes other types of stopwords such as contractions.[4] Most languages have lists of common "stop words" that can be provided to pre-processing programs we discuss below.

We note that for every language choosing which stop words should be removed is a substantive decision that in some cases can have important effects on the results of the analysis. For example, Campbell and Pennebaker (2003) study the importance of pronouns which could be considered stop words in some schemes. Fokkens et al. (2013) find that differing removal of stopwords can produce different results in some cases. In other words, choosing a stopword list should be carefully chosen, based on words that the analyst thinks will not be important in informing the analysis. We discuss how we use stopwords in more detail in the specific applications (both multi-lingual and single language) below.

Stemming and Lemmatization Stemming removes the endings of conjugated verbs or plural nouns, leaving just the "stem," which in many languages is common to all forms of the word. Stemming is useful in any language that changes the end of the word in order to convey a tense or number, which includes English, Spanish, Slovenian, French, modern Greek and Swedish. Since tense and number are generally not indicative of the topic of the text combining these terms can be useful for reducing the dimension of the input. However, not all languages require stemming. For example, Chinese verbs are not conjugated and nouns in Chinese are usually not pluralized by adding an ending. A host of studies have shown stemming to be an effective form of preprocessing in English, however the benefits are both application and language specific(Salton, 1989; Harman,

[3]See Online Appendix A for additional discussion.

[4]In other settings, such as the analysis of style or authorship detection, function words may be the sole quantity of interest (Mosteller and Wallace, 1963).

1991; Krovetz, 1995; Hull, 1996; Hollink et al., 2004; Manning et al., 2008).[5]

Stemming is an approximation to a more general goal called lemmatization – identifying the base form of a word and grouping these words together. However, instead of chopping off the end of a word, lemmatization is a more complicated algorithm that identifies the origin of the word, only returning the *lemma*, or common form of the word. Lemmatization can also determine the context of the word, for example it will leave *saw* the noun as is, but will turn *saw* the verb into *see* (Manning et al., 2008). While stemming often works almost as well as lemmatization in languages like English, lemmatization works better for languages where conjugations are not indicated by changing the end of the word, and for agglutinative languages[6] where there is a greater variety of forms for each individual word, such as Korean, Turkish, and Hungarian.

Compound words Some languages will frequently concatenate two words that describe two different concepts, or split one word that describes one concept. These instances, called compound words or decompounded words, can decrease the efficacy of text analysis techniques because one concept can be hidden in many unique words, or one concept may be split across two words. For example, the German word "Kirch," or church, can be appended to "rat," forming "Kirchrat" who is a member of the church council, or "pfleger" to form "Kirchenpfleger," or church warden. If it is appended, the computer will not see "Kirch" as an individual concept. Decompounding this case would separate "Kirch" from its endings. "Compounding languages" include German, Finnish, Danish, Dutch, Norwegian, Swedish, and Greek (Alfonseca et al., 2008). On the other hand, the analyst may want to compound words. For example, in English "national security" and "social security" each contain two separate terms even though they express one concept. Even though they share the word "security", these concepts are very different from each other, so the analyst might wish to compound these into "nationalsecurity" and "socialsecurity."

[5]Several computer programs are available to implement stemming, including `txtorg` (discussed in Section 2.3), which can implement stemming in multiple different languages. These programs automatically detect common variations in word endings, removing these endings, and plural words into their singular form.

[6]Languages where most words are formed by combining smaller meaningful language units called morphemes.

All of these decisions should be guided by substantive knowledge.

Segmentation Some languages, like Chinese, Japanese and Lao, do not have spaces between words and therefore text analysis techniques that rely on the word as the unit of analysis cannot naturally parse the words into individual units. Automatic segmentation must be used before the documents can be processed by a statistical program (see Lunde (2009) for an overview). Segmentation can be done using dictionary methods (Cheng et al., 1999) or using statistical methods that learn where spaces are likely to occur between words (Tseng et al., 2005).

2.2.3 Building the Document-term Matrix

Once all pre-processing has been completed, for many automated content techniques (including those detailed in this paper), the remaining words are used to construct a document-term matrix (DTM). A document-term matrix is a matrix where each row represents a document and each column represents a unique word. Each cell in the matrix denotes the number of times the word indicated by the column appears in the document indicated by the row. For example, if a document was just the sentence "I support the Tories", "I" and "the" would likely have already been removed as stop words, so that the document would be represented with a 1 for "Tories" and "support" and a 0 for all other words.

Following the "bag of words" assumption the DTM format preserves information about how many times each word appears in a document while discarding information about the word order. The resulting matrix is extremely sparse, meaning a large proportion of the cells are zeroes, because most documents will contain only a small fraction of the words in the vocabulary. For even moderately sized corpora this matrix will be too large to store in its rectangular form; however, we can exploit the sparsity of the DTM to store only the non-zero entries. The DTM, or its sparse representation, is the primary input to most automated text analysis methods including the ones in this article.

2.3 Multi-language pre-processing tools

2.3.1 Language-specific processing

All of the previous steps are not trivial from a workflow perspective, especially for comparativists working in a variety of languages, each of which may require specialized tools. Here we discuss existing methods to deal with pre-processing text within a language. There are two flexible open source software tools for doing stemming, stopword removal, etc., that cover many languages. First is the the R package `tm` (Feinerer et al., 2008), which can stem 11 languages.[7] and can do stop word removal on 13 languages[8] Another tool is the Python/Lucene-based application `txtorg`[9], which currently includes support for 32 languages[10]. In `txtorg`, all supported languages go through a suite of best practice preprocessing steps, which includes the appropriate combination of stemming, segmentation, and stopword removal for that particular language. Both of these tools facilitate text pre-processing, though `txtorg` is dramatically more efficient in handling larger corpora and when searching and subsetting large amounts of text.[11]

2.3.2 Translation

As we discuss in Section 3.2 and illustrate in Section 4.2, there are important instances where modeling textual data from multilingual corpora becomes more efficient and accessible for applied users if the text is first translated into a single language. Of course, though human translation remains the gold standard, the scale of textual data generally far exceeds that which might be feasibly translated by humans. In subsequent sections we discuss the relevant technical considerations of multi-lingual analysis in greater detail.

[7]Danish, Dutch, English, Finnish, French, German, Norwegian, Portuguese, Russian, Spanish, Swedish.

[8]Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Russian, Spanish, and Swedish.

[9]Note that `txtorg` includes a graphical user interface built with `TkInter`, so users do not need to know Python in order to use `txtorg`. Nearly all `txtorg` functionality is accessible without writing any code.

[10]Arabic, Armenian, Basque, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, English, Finnish, French, Galician, German, Greek, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Korean, Latvian, Norwegian, Persian, Portuguese (separate tools for Brazil and Portugal), Romanian, Russian, Spanish, Swedish, Thai, and Turkish.

[11]See Online Appendix G for some basic benchmarking information between `tm` and `txtorg`.

In this section, we briefly discuss machine translation and introduce an `R` based utility for accessing machine translation software developed by Google and Microsoft.

Central to comparative politics is, of course, a commitment to cross-national comparison. And while comparativists have developed many techniques for automated text analysis, there presently exists little or no support for cross-lingual comparison. While this limitation does not preclude all potentially interesting comparisons, it prevents a great many. In Section 3.3, we discuss a principled way by which such comparisons can be made with the Structural Topic Model after first translating the corpus into a common language. However, this requires first overcoming the potentially formidable task of translating the data into a common language.

The job of a translator is to "render in one language the meaning expressed by a passage of text in another language" (Brown et al., 1990, p.81), and though there exist many approaches, the basic task of machine translation is to accomplish this conversion with a computer. Because of its many uses and because early barriers to machine translation, which included hardware limitations and a dearth of machine-readable text (Brown et al., 1993), have largely been overcome, there is now heavy investment in machine translation. There exist a number of academic and commercial labs committed to the development of machine translation systems, some of which have led to the founding of new companies. Simultaneously, large, mature software companies like IBM, Microsoft, and Google have also developed their own machine translation systems (Koehn, 2009). Because these groups can leverage financial and academic resources beyond those generally accessible to political scientists, we argue that a desirable solution to the problem of machine translating text for political science is one that leverages the effort made by dedicated research groups in a simple, straightforward way.

When translating text for eventual consumption by human readers, there can be no substitute for human translation. Within the literature on machine translation evaluation, it is said that "The closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2002, p.1). But compared to translating text for eventual consumption by human readers, translation for multi-lingual text analysis is a slightly

easier problem. As discussed in Section 2.2, most approaches to automated text analysis make a bag-of-words assumption, which implies that the ordering of terms in a document does not matter. The translation software need only correctly translate the significant terms in the original document, as any error in word order will be discarded by the bag-of-words assumption.

If users want to use machine translation, what should they use? Our answer is to provide an `R` package, `translateR`, that permits easy access to two very mature translation systems, namely those produced by Google and Microsoft. The package supports a variety of input and output formats and can be easily used with other text analysis software. Crucially for our purposes, the package preserves information about individual texts (such as the original language or date of authorship). This is important for using models like the Structural Topic Model that incorporate this data. Moreover, `translateR` preserves the scalability of machine translation by the translation process via multiple API calls. Users provide as input the data to be translated, either as a dataframe with metadata or as a vector of documents or terms, and `translateR` makes calls in parallel to the translation API specified by the user (either Bing or Google). As a result, researchers spend minimal time reformatting their data and similarly little time waiting for the translation process to finish along with other aspects necessary for standard textual analysis. Additional discussion and syntax is given in Online Appendix C.

3 Computer Assisted Text Analysis

In the previous section we discussed in detail how to prepare a multilingual corpus for automated approaches to text analysis by creating a document-term matrix. A complete overview of methods for quantitatively analyzing the text is beyond the scope of this paper. Unlike the issues involved in multilingual text processing, these methods have been well developed elsewhere (e.g., Grimmer and Stewart, 2013). In the next section (3.1), we provide a brief, selective overview and direct interested readers to our online appendix which provides an accessible introduction to a broader range of methods. We then discuss the challenges that arise in moving from single to multilingual corpora (Sec-

tion 3.2). Finally in Section 3.3 we describe the Structural Topic Model before providing two applications of its use (Section 4).

3.1 A Brief Overview of Approaches

There are essentially two approaches to automated text analysis: supervised and unsupervised methods, each of which *amplifies* human effort in a different way. In *supervised* methods we specify what is conceptually interesting about documents in advance, and then the model seeks to extend our insights to a larger population of unseen documents. Thus for example, we might manually classify 100 documents into two categories with the model classifying the remaining 9900 documents in the corpus. In *unsupervised* methods, such as topic modeling, we do not specify the conceptual structure of the texts beforehand. Instead we use the model to find a low-dimensional summary that best explains observed documents given some set of assumptions. Consequently human effort shifts from construction of a training set in supervised learning to interpretation of the model results in unsupervised settings.

In our applications we leverage a particular type of unsupervised topic modeling built on the popular Latent Dirichlet Allocation (LDA) model (Blei, 2012). LDA is a mixed-membership model which means that each document is represented as a mixture over a set of topics and each observed word is conditionally independent given its topic.[12] Each topic is a distribution over the words in the vocabulary which crucially are learned rather than assumed by the model. LDA has seen widespread use in computer science and the humanities due to its simple and extensible structure.

The full range of text analysis methods including supervised and unsupervised methods are discussed in greater detail in Grimmer and Stewart (2013). We have also included an online appendix for this paper containing an abbreviated introduction using a consistent set of heuristic examples using data from a corpus of comparative politics papers published in the *American Political Science Review* (Online Appendix B).

[12]By 'mixture' in this context we mean a set of positive values that sum to one.

3.2 Multilingual text modeling

A considerable advantage to the quantitative approach to text analysis is that the methods are language agnostic. However, a rarely discussed limitation is that the documents are assumed to be drawn from only one language. This can be a frustrating situation for practitioners in comparative politics who are interested in studying a multilingual corpus. Here we discuss the attendant methodological issues that apply to both supervised and unsupervised models.

In some respects, the most natural approach for handling a multilingual corpus is to perform analysis within the native language but referencing a commonly shared objective. This is the approach taken in manual coding efforts, such as the Comparative Manifestos Project (Volkens et al., 2013), where it is relatively straightforward to define the coding criteria in a language independent way but analyze each document in its own native language. For keyword and supervised approaches, it is plausible to develop a separate but statistically comparable dictionary or training set for each observed language. Unlike the manual case where a single codebook can be developed in a shared language, the automated approaches require a duplication of effort for each language. While feasible in supervised settings, there isn't a a clear analog for unsupervised methods.

A second approach is to translate text into a common language. Manual translation by an experienced translator would be extremely costly and so we turn to machine translation tools introduced above. How well this works will depend on the quality of the machine translation and the goal of the analysis. We return to this approach in Section 4.2.

The third approach is to develop a model which maintains an explicitly multilingual representation. The central challenge is to develop an alignment between the conceptual representations of the model across languages so that we know a particular scaling, topic or class in one language is comparable with the representation in another language. We focus here on the challenging case of unsupervised topic models in the style of LDA where the conceptual representation is being learned from the data, although the general ideas apply straightforwardly to supervised methods as well.

Existing approaches to multilingual topic models are differentiated in how they lever-

age external information to implicitly or explicitly align comparable topics across languages. The Polylingual Topic Model of Mimno et al. (2009) leverages a set of aligned documents, for example Wikipedia articles on the same topic in different languages. By constraining aligned documents to share a distribution over topics, the model is able to align the words associated with a given topic across languages. The Bilingual Topical Admixture model (Zhao and Xing, 2006) works with texts which are aligned at the token level (such as through the result of machine translation). Exact translations which are aligned at the token level are more difficult to obtain, but they provide a more direct source of information about topic alignment. Finally the Multilingual Supervised LDA model (Boyd-Graber and Resnik, 2010) uses a combination of sentiment information and aligned dictionaries to develop multilingual topics. Recent work combines these approaches to leverage both dictionary and document level alignments simultaneously resulting in a model which is more robust than either independently (Hu et al., 2014).[13]

From a technical perspective fitting most of these models involves a relatively straightforward adaptation of the collapsed Gibbs sampling algorithm for LDA (Griffiths and Steyvers, 2004).[14] The result is a set of topics for each language along with the document-topic loadings. Multilingual models have primarily been used for either document exploration or machine translation tasks.

The existing models for multilingual analysis do, of course, have limitations. The correspondence between the multilingual topics relies on the particular alignment information provided by the user and needs to be validated. This can be particularly challenging for indirect strategies such as the document alignment in the Polylingual Topic Model. For each topic the user needs to verify that the topic-word distributions are comparable across languages. Given that the size of the vocabulary may be in the thousands, assessing model

[13]Boyd-Graber and Blei (2009) introduce a topic model for completely unaligned texts, but they note that the model is highly sensitive to starting values and when run to divergence can result in the nominally equivalent topics between languages diverging. This is evidence for the central role of observed alignment information in pinning down the correspondence between topics.

[14]For example, for the Polylingual Topic Model, we iteratively sample each token in the document adjusting the topic-word distribution for the language specific version of the topic but sharing the document-topic counts across all languages within the document. This algorithm has comparable speed to LDA but with slightly higher memory requirements.

failure can be a substantial challenge even for only two languages.[15] While the articles described above provide diagnostic tools for the model results, they are primarily focused on the machine translation applications that motivate that literature.

As a practical matter, multilingual topic models generally lack the ability to include additional document metadata, which we argue below is an important part of applied social science research. In addition, there are limited software tools available for the estimation of these models.[16] These critiques are not problematic for the models as presented in their original context, but do suggest challenges for their use in applied comparative research. Below we suggest a way that machine translations and the Structural Topic Model can be fruitfully combined.

3.3 The Structural Topic Model

In our applications (Section 4) we leverage a recently introduced framework, the Structural Topic Model (STM) (Roberts et al., 2014, 2013, nd). The STM is a mixed-membership topic model (like Latent Dirichlet Allocation (LDA)) with extensions that facilitate the inclusion of document-level metadata.[17] The inclusion of this information within the model can both improve the quality of the learned topics and facilitate hypothesis testing. Software for estimating the model is freely available in the `R` package `stm`.

Before moving on to our applications of STM, we first briefly review several aspects of our use of the STM which are specific to this context. A brief statement of the model is available in Online Appendix D. For additional technical details on estimation and implementation of the model we refer to existing work (e.g., Roberts et al., 2014, nd, 2013).

[15] As the number of languages grows this problem is compounded by the need to have a single scholar who reads all languages. For example, amongst our team no author speaks both Arabic and Chinese which would make direct validation of a Polylingual Topic Model quite difficult.

[16] Of the models discussed here, only the Polylingual Topic Model of Mimno et al. (2009) has a publicly available software implementation. A Java implementation is available in the software package Mallet (McCallum, 2002).

[17] The inclusion of document metadata follows and extends two developments within political science. The Dynamic Topic Model (Quinn et al., 2010) is a single membership model in which the probability of observing a topic moves smoothly through time. The Expressed Agenda Model (Grimmer, 2010) is a single membership model which includes information about document authors. However no such model exists to include author and time simultaneously. Drawing on these works, our approach generalizes to arbitrary covariate information and extends these setups for the mixed-membership case.

The Role of Covariates STM differs from other topic modeling techniques like LDA in allowing document-level covariates to be included in the model as a method for pooling information. A covariate can be allowed to affect either *topical prevalence* or *topical content*. Covariates in topical prevalence allow documents to share information about which topics are expressed within the document (e.g. women are more likely to talk about topic 1 than men). Users can plot the relationship between their topic prevalence covariates the expected proportion of a document that belongs to each topic. Covariates in topical content allow for the rates of word use, for each topic, to differ by covariate values (e.g. women are more likely to use a particular word when talking about a particular topic than men). Users can include both prevalence and content covariates, only one type, or neither.

Content covariates are a particularly powerful tool which can be used to capture both quantities of interest and condition away systematic differences within the corpus that are not of primary interest. Imagine for example we were attempting to compare topical coverage within a large corpus of news reports about China from Agence France Presse (AFP) and Xinhua, China's state news agency. In order to facilitate a direct comparison we want the model to discover (for example) a single topic on Tibet; however, systematic differences in the way that AFP and Xinhua cover Tibet may produce separate AFP-Tibet and Xinhua-Tibet topics. Instead by allowing the model to maintain an AFP version of the topic and a Xinhua version of the topic (which are constrained to be close), we can estimate the differences in word use and still retain a straightforward comparison. If the differences are themselves of interest, the analyst can compare words distinctive to Xinhua version of the topic ("oil", "gas", "resources") with words distinctive of the AFP version ("culture", "religion", "independence").[18] If the differences are simply a nuisance we can marginalize over source-specific version of the topics weighting by the document frequency within the corpus as a whole. We will return to this idea in our multilingual analysis where we use content covariates to condition out systematic differences that result from translation to English from different languages.

[18]The example here is drawn from the data and model described in (Roberts et al., nd).

Topic Correlations In addition to the inclusion of covariates, the second distinctive feature of the STM is the explicit estimation of correlation between topics.[19] Graphical depictions of the correlation between topics provide insight into the organizational structure at the corpus level. In essence the model identifies when two topics are likely to co-occur within a document (here we focus on positive correlations although negative correlations are also estimated). The software we provide allows the user to produce a network graph of topics where each topic is a node and two nodes are connected when they are highly likely to co-occur. This can help the user to identify larger themes that transcend topics.

Drawing on recent literature in undirected graphical model estimation, we extend the approach developed in Blei and Lafferty (2007) for estimating the edges of the graph. In Online Appendix E we describe the two graph estimation procedures we provide along with parameters set by the user. We give a specific example of this approach in the next section.

4 Applications

In this section we introduce two applications of the STM. The first application, the analysis of Islamic fatwas is conducted entirely within the single native language. For the second application, our corpus includes both Chinese and Arabic texts which we translate into a common language prior to analysis. All the analysis tools used below are built in to the R package `stm` (Roberts et al., 2014).

4.1 Jihadi Fatwas

In this example, we combine data on Muslim clerics from Nielsen (2013) with expert coding of whether clerics are Jihadist or not to see how the topical content of contemporary Jihadist religious texts differ from those of non-Jihadists. Nielsen collects data on the lives and writings of 101 prominent Jihadist and non-Jihadist Muslim clerics, including the 27,248 texts available from these authors from online sources. A majority of these

[19]Correlations are estimated by replacing the Dirichlet distribution in the standard LDA framework with a logistic normal distribution as in the Correlated Topic Model (Blei and Lafferty, 2007). When no covariates are specified, the STM reduces to an instance of the Correlated Topic Model.

texts are *fatwas* – Islamic legal rulings on virtually any aspect of human behavior, ranging from sex and dietary restrictions to violent Jihad. For many clerics, Nielsen also collects books, articles, and sermons on the same types of topics. Collectively, these texts are representative of how clerics choose to interact with religious constituencies; in fact many of these collections are curated by the clerics themselves.

We combine these texts with an independent coding of whether these clerics are Jihadist or not based on two scholarly sources. First, the *Militant Ideology Atlas - Executive Report* (McCants, 2006), Appendix 2, lists 56 individuals that are frequently cited by Jihadists. The authors of the Atlas code whether these are "Jihadi authors" according to substantive knowledge. Second, Jarret Brachman (Brachman, 2009, pp. 26-41) lists the names of prominent clerics in eight ideological categories: establishment Salafists, Madkhali Salafists, Albani Salafists, scientific Salafists, Salafist Ikhwan, Sururis, Qutubis, and Global Jihadists. The latter two categories are Jihadist while the rest are not. These two sources largely overlap; together, they provide expert assessments 33 of the clerics (20 Jihadists and 13 non-Jihadists) for whom Nielsen collects 11,045 texts.

We then estimate a Structural Topic Model with the binary indicator for Jihadi status as a predictor. The results are shown in Figure 1, with topics presented as collections of words (in this figure, we leave the words in Arabic), along with the topic coefficients and standard errors. We estimate 15 topics after experimenting with 5- and 10- topic models that produced less readily interpretable topics.[20]

The first inferential task is to infer topic labels from the words that are most representative of each topic. We do this by examining the most frequent words in each topic and the words that have the highest levels of joint frequency and exclusivity (meaning they are common in one topic and rare in others. In several cases, we also examine *exemplar documents* for a topic — those documents that have the highest proportion of words drawn from the topic. This also serves as a validation step because we check whether words in the topic have the meanings in context that they appear to have in the topic

[20]This is not to say that 15 is the "right" number of topics in this corpus — rather we find a 15 topic model for uncovering useful insights about the structure of the texts in relation to the Jihadist ideology of their authors.

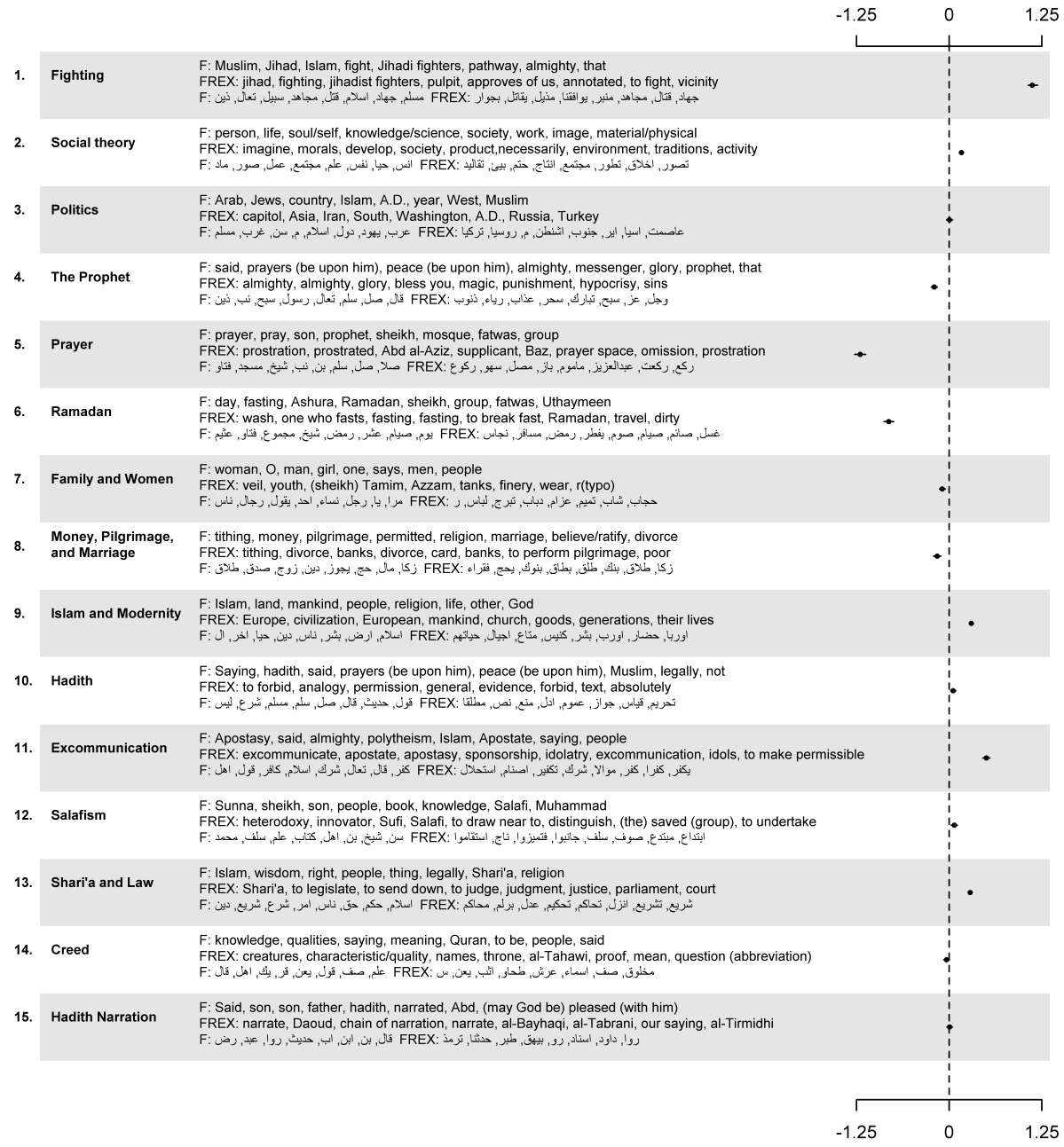

Figure 1: Coefficients and standard errors for a 15-topic Structural Topic Model with Jihadi/not-Jihadi as the predictor of topics in Arab Muslim cleric writings. The words used to label each topic are shown on the left. "F:" indicates words that are most frequent in each topic. "FREX:" indicates words that are frequent *and* exclusive to each topic. The Arabic words are in their stemmed form.

frequency lists.

The results in Figure 1 indicate that topics 1 (Fighting) and 11 (Excommunication)

are most correlated with the indicator for Jihadist clerics, matching our *a priori* predictions based on the content of the topics. Excommunication (*takfīr* in Arabic) is commonly used by Jihadists to condemn fellow Muslims who disagree with Jihadist aims or tactics. The exemplar documents for this topic are fatwas on the rules and justifications for excommunication and other writings that make heavy use of the concept of excommunication. In contrast, topic 1 is a broader Jihadist topic focused primarily on fighting the West — the exemplar documents are fatwas about fighting abroad. Topics on social theory, Islam and modernity, and Shari'a and law are also correlated with Jihadism, though to a lesser degree.

A number of other topics are also clearly identifiable, including topic 5 on prayer, topic 6 on Ramadan, and topic 8 on money, pilgrimage, and marriage. As we expected from their content, these topics receive relatively little attention from Jihadists who are more focused on their violent struggle than with fine distinctions in Islamic legal doctrine and religious ritual.

We can use the estimated correlation of topics with other topics to learn more about the structure of the corpus.[21] In Figure 2, we plot the network of topics such that topics that are correlated are linked. Many of the correlations between topics are intuitive and revealing about the nature of Islamic legal discourse. The topic on *hadith* (the sayings of the Prophet Muhammad) is highly correlated with language about the chain narration by which each *hadith* is verified as trustworthy. Authors who write about social theory are likely to also write about Islam and modernity, politics, the role of women, and Shari'a and law.

Figure 2 shows correlations between topics preferred by Jihadists. Documents that include language about excommunication tend to also include text about creed (what Muslims believe), shari'a and law, the Prophet, and fighting. Documents about fighting are likely to also include politics, discussions of Salafism, Islam and modernity, and Shari'a and law. In contrast, texts about non-Jihadi legal issues — prayer, Ramadan, Money, Pligrimage, and Marriage — are unlikely to be about more than one topic. This aligns

[21] We introduce our approach to calculating and graphically representing the correlation structure in Online Appendix E.

with our qualitative assessment of the corpus: the modal Jihadist fatwa is article length and ranges across multiple topics while the modal non-Jihadist fatwa is paragraph-length and gives a precise ruling on only one topic.

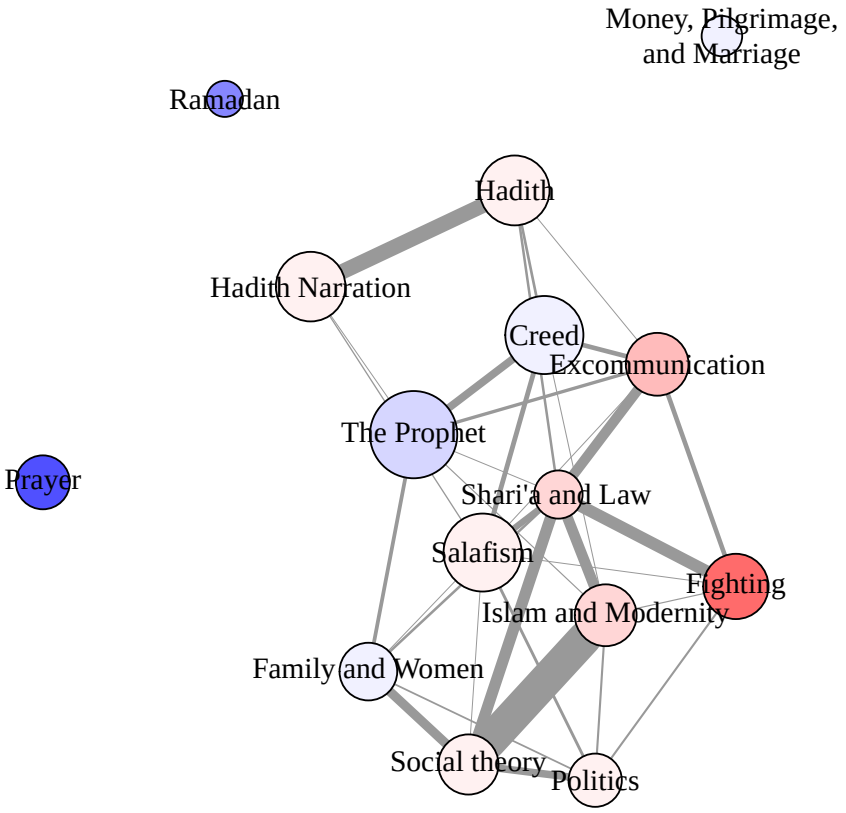

Figure 2: The network of correlated topics for a 15-topic Structural Topic Model with Jihadi/not-Jihadi as the predictor of topics in Arab Muslim cleric writings. Node size is proportional to the number of words in the corpus devoted to each topic. Node color indicates the magnitide of the coefficient, with redder nodes having more positive coefficients for the Jihadi indicator and bluer nodes having more negative coefficients. Edge width is proportional to the strength of the correlation between topics.

The presence of at least two clearly Jihadist topics invites further inquiry. Figure 2 shows that these topics are correlated in general, but do all Jihadists write on both topics? Do some write more on one? Does this split indicate an intellectual divide within the Jihadist subgroup? To take a first cut at these questions, we simply plot the proportion of the *Excommunication* topic against the *Fighting* topic, as shown in Figure 3. The results teach us several new things about how Jihadists and non-Jihadists write. First, for many Jihadists, document space spent on *Fighting* is substitute for space spent on

Excommunication.[22] Usama bin Laden has the highest proportion of words devoted to *Fighting* — about 38 percent — but he spends only two percent of his words discussing the excommunication topic. This accords with Bin Laden's long-time focus on the goal of targeting and provoking the West through both writings and deed.

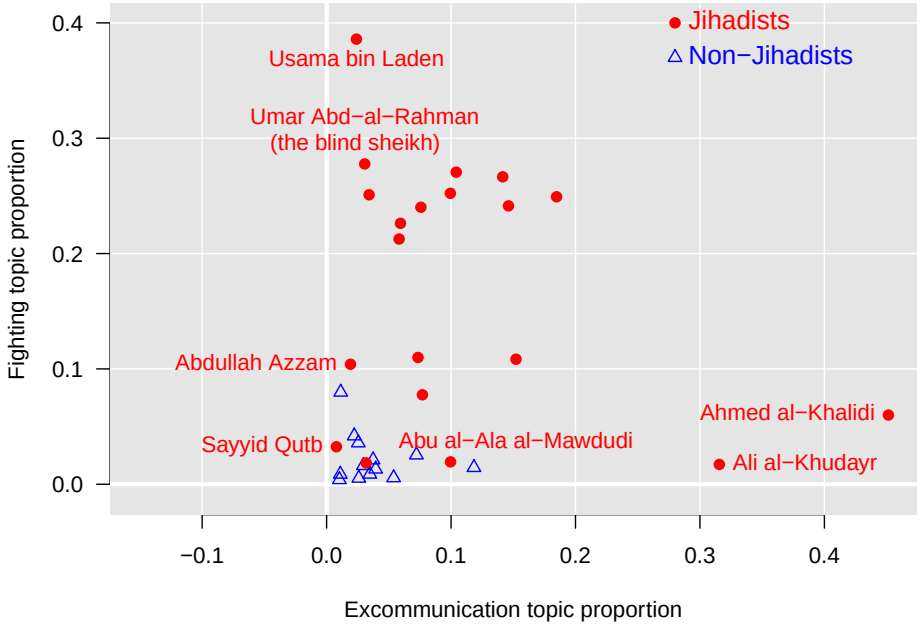

Figure 3: Estimated topic proportions by fighting the west and excommunication topics, separated out by jihadist versus jihadist coding.

At the other extreme, Ahmad al-Khalidi and Ali Khudayr respectively devote 46 and 32 percent of their writing to excommunication and almost none to fighting. This is not surprising when we consider the life-trajectories of these clerics. All both have issued fatwas excommunicating prominent Muslims for alleged heresies and both have spent time in Saudi prisons for doing so. This finding adds further face validity to our findings — the clerics most interested in writing about excommunication of fellow Muslims are those that have also carried it out repeatedly. Between these endpoints, most other Jihadists spread out on a continuum where more discussion of excommunication means less of fighting and vice versa. It is likely that these two topics are virtually all that some of these authors

[22]This is not inconsistent with the finding that these two topics are correlated within texts. The presence of one topic increases the likelihood of the presence of the other topic in a text, but some authors focus on one topic more than the other.

write about. Given that filler words and others must still be assigned to topics, it may simply be the case that no more than 50 percent of a document can be allocated across these Jihadi topics.

Several Jihadist authors have low enough proportions of both Jihadist topics that they could be mistaken for non-Jihadist clerics. Sayyid Qutb is often considered one of the founders of the Modern worldwide Jihadist movement, but only three percent of his writing is devoted to the topics that tend to occupy other Jihadists. Similarly, Abu al-Ala' al-Mawdudi and Abdullah Azzam are considered canonical authors by Jihadists, but only about 10 percent of their writings are devoted to the topics of *Fighting* and *Excommunication*.

To see what is unique about the writing of these authors, we look at the topics to which they devote the most attention and find that their profiles are very similar. Each devotes the bulk of their writing to writing about social theory, politics, and Islam and modernity. We find that the current Jihadist focus on fighting the West and excommunication is relatively new. We show this in Figure 4 by summing the proportion of writing that each Jihadist author devotes to either excommunication or fighting and plotting it against the year of each cleric's birth. Among the set of individuals identified in the secondary literature as Jihadists, only those of relatively recent vintage are writing on the two topics that are now core to Jihadist ideology.

To summarize, we find that a 15-topic model provides insight into the structure of an Islamic legal corpus that includes work by Jihadists and non-Jihadists. Although one might expect Jihadism to be monolithic, there are in fact multiple ways that Jihadists write about their subject. In particular, there is suggestive evidence of a trade-off for many Jihadists between focusing on fighting the West and focusing on excommunicating fellow Muslims they feel are inadequately supporting the Jihadist cause. We also find that an older generation of Jihadist writers does not write about either of these topics, suggesting that Jihadist writing was more eclectic in the past but has become homogenized over time.

Figure 4: The proportion of words by each Jihadi author devoted to excommunication or fighting, plotted against the year of their birth with a best-fit line.

4.2 Reactions to Snowden in China and the Middle East

In this section, we provide an illustrative example of how machine translation can be used in conjunction with the Structural Topic Model to make comparisons across countries and languages. An important theoretical and empirical agenda is understanding how other countries view the United States (Katzenstein and Keohane, 2007; Chiozza, 2009; Lynch, 2007; Telhami, 2002; Rubin, 2002). One way to understand views of the US is to compare responses to specific events (e.g., Jamal et al., nd). Here we look at responses to a single event across different language communities. We collected thousands of social media posts in Arabic and Chinese during June of 2013, the month when former U.S. government employee Edward Snowden disclosed thousands of classified documents that detailed the U.S. government's clandestine surveillance program. Because the documents leaked by Snowden contained many revelations about surveillance of and cooperation with other countries, some scholars have worried that the leaks would undermine U.S. legitimacy abroad (Farrell and Finnemore, 2013). We focus on the reaction of citizens in China and the Middle East, arguably two of the most important U.S. strategic areas in the world.

We generate our corpus by collecting the universe of unique posts from Twitter in

Arabic containing the word for "Snowden" and the universe of unique posts from Sina Weibo containing the word for "Snowden" in Chinese from June 1-June 30, 2013.[23] Twitter is banned in China, so collection of Twitter posts in Chinese would contain those of foreign Chinese-speakers, or of those who are sophisticated enough to jump the Great Firewall, and therefore would be a potentially biased sample. Sina Weibo is the closest comparable platform to Twitter in China.[24]

4.2.1 Two Approaches to Machine Translation

Ideally we want to analyze both Arabic and Chinese within the same topic model. Leaving the two corpora in their respective languages would lead to essentially no overlap in vocabulary between the Arabic and Chinese posts. As a result, each corpus would have their own individual topics, since the model cannot recognize that Snowden in Arabic is the same word as Snowden in Chinese, rendering direct comparison of topical content essentially impossible. As described in Section 3.2, we need to use some type of external alignment between between languages to analyze the two corpora within the same model. Translation provides alignment by creating overlap between the two corpora. Here we explore solutions based on machine translation as software implementations are widely available and continuously improving. We use two approaches to machine translation: translating the entire corpus and translating only terms that appear in the document-term matrix. Both approaches easily extend to document sets containing more than two languages.

In the first approach, we use machine translation to translate both corpora of text completely into a common language, English. There are compelling reasons to translate to a "third-party" language, particularly when that language is English. Perhaps the most basic reasons for choosing English is the ability to communicate research findings to an English speaking audience. We also wanted both sets of text to undergo the same amount of translation. If we had translated the Arabic into Chinese, for example, and

[23] We point readers interested in the preprocessing that we conducted on the Snowden corpus to Online Appendix F.3.

[24] Both Weibo and Twitter restrict the number of words within posts. All data obtained from Crimson Hexagon.

left the Chinese text untranslated, the Chinese corpus might dominate to topic model as it would have no words that were "untranslatable." This at least makes it more plausible that the inevitable error introduced in translation is roughly comparable between the two language groups, resulting in a type of symmetry. Of course this may not be the case when two languages are more closely related or where the translation accuracy is substantially higher for one kind of transformation.

Beyond the appeal of symmetry, English is a particularly useful common language due to its role in machine translation systems. Most modern machine translation systems use parallel corpora to learn the parameters of a statistical model. However, many language pairs don't have large parallel corpora easily available and so instead a "pivot", or bridge, language is used as an intermediate point in creating the translation. English is a common pivot language due to the widespread availability of texts. Thus not only would we expect the Chinese to English and Arabic to English to have particularly high accuracy, but for many machine translation systems a translation between Chinese and Arabic will involve a translation through English.[25] For more on pivot languages in statistical machine translation systems we refer readers to Utiyama and Isahara (2007) and Paul et al. (2009). Habash and Hu (2009) discusses the specific case of using English as a pivot language for Chinese and Arabic.[26] We use Google Translate to perform translation, passing each post through `translateR` and recording the translation. Online Appendix F discusses our pre-processing steps.

The complete corpus strategy is ideal because it introduces no additional sources of information loss beyond the machine translation process. Because each original text is translated, words are always considered within the context that they appear. Con-

[25] As a proprietary system we do not know for sure if Google Translate uses English as a pivot language for Chinese and Arabic. However even if it does not, we can expect that it would provide reasonable results based on the widespread availability of English parallel corpora (e.g. Linguistic Data Consortium catalog).

[26] For researchers looking to apply these methods to their own texts we recommend English as a useful default choice for a common language, even if some of the documents are already in English. In particular circumstances with language groups which are closely related or where excellent parallel text corpora or available a different common language may be more appropriate. The applied researcher can always investigate different options by informally evaluating translation quality by using Google Translate to process a small sample of documents.

text not only improves accuracy in most machine translation system, but may, in some cases, be necessary for an appropriate translation. The downside is that the process of machine-translating a corpus of even a few thousand documents can be expensive and time-consuming because all the text is passed to the machine translation service.

Given these considerations, we also investigate a second approach which relies on only the minimal number of translation queries. We first created a document-term matrix for each language's corpus separately and translate only those terms. We take the intersection of the two translated vocabularies and merge the document-term matrices together. While this approach discards word context within translation, it is considerably cheaper.[27] The cost of translation for the complete corpus grows linearly with the size of the corpus because every occurrence of every unique term is translated. By contrast, in the term by term translation, the marginal cost of translating an additional document decreases as the corpus grows, because there are fewer and fewer unique terms in each additional document as more documents are added to the corpus.

In our case, the two approaches give somewhat comparable results. We *strongly* caution though that this may not be true in general. Fortunately validation of the translation strategy is relatively straightforward. The natural first check is to verify that topics are not exclusively related to a particular language.[28] If this is not a problem, reading documents highly associated with a particular topic in the native language provides a validation of the translation process. If the documents are largely in agreement with the semantic meaning of the concept as represented in the new language, then the loss of

[27]For our corpus, the full document translation costs approximately $450, whereas the term translation was approximately $10 (both with Google Translate accessed through `translateR`). In general, as of summer 2014, translation with the Google API costs $20 per 1 million characters of text, so 500,000 characters costs $10, 2 million characters costs $40, etc (more information at `https://cloud.google.com/translate/v2/pricing`). The Microsoft Translator API operates with a very different cost structure. Users sign up for a monthly plan, which caps the total number of characters that can be translated in a single month. It is free to translate up to 2,000,000 characters per month, $40 for 4,000,000 characters per month, $160 for 16,000,000 characters per month, etc (more information at `https://datamarket.azure.com/dataset/bing/microsofttranslator`). Note that for both Google and Microsoft, a "character" means an escaped, url-safe character, so documents written in a language like Chinese often become three to four times longer. However, `translateR` automatically converts the characters to their url-escaped versions, so users do not need to do so manually.

[28]Note that this need not signal a problem as a topic could actual be specific to a particular country or language. We merely include this to emphasize that such findings should be checked to ensure that they didn't arise by a failure in the translation process.

information from the approximate translation procedure is likely acceptable. Analysts should of course be attentive to the way that systematic errors in translation will affect the particular argument that they wish to make and adjust accordingly.

While the complete corpus translation approach is to be preferred in general, the term translation strategy can provide a cost effective alternative in particular cases. We imagine that this might be particularly useful for early exploratory analyses which can be used to justify the greater expenditure of complete corpus translation approaches.

4.2.2 Correcting for Systematic Differences Between Languages

As discussed in Section 2.3.2, machine translation is not an error free process. In either of the approaches discussed above, there will be untranslated words, mistranslated words, or words with multiple meanings. As such, words that mean the same thing in the Chinese and Arabic corpus could sometimes map onto different words in English that are synonyms of each other. Just as a native Arabic-speaker would speak English differently than a native Chinese-speaker, using a vocabulary and sentence structure that most closely maps onto their respective native languages, the 'way' in which machine translation interprets each language will be different for the two different corpora. These linguistic differences pose a challenge for topic models. We want to ensure that the topics are uncovering differences in semantic content rather than linguistic idiosyncrasies in describing that content.

As discussed in Section 3.3, the STM allows for this facet of a corpus. Within the STM we can use a *content* covariate to capture variations in word use attributable to observed covariates. Here we include the document's original language as a content covariate in order to capture linguistic differences in describing a topic. This allows us to effectively marginalize over differences in word rate use that arise due to linguistic differences or errors in translation. For example, the Chinese word for liquor translates into "wine" in Google translate. The Arabic word for liquor, however, translates into "spirits". If there were a "party" topic within our corpus, this would allow both the Chinese and Arabic documents to talk about the party, but the Chinese version of the translation would use wine slightly more and the Arabic translation would use spirits slightly more. Crucially,

there are a set of common words that do overlap between the two languages, which allow us to learn that these systematic differences between the languages are related words and not completely separate concepts.[29]

4.2.3 Results

Next we discuss the results of our illustrative analysis. For all of our analyses, we used a 15 topic model, using an indicator variable for what language community generated the social media post as both a topic prevalence and content covariate, as well as a smooth function of time (date of the post) as a topic prevalence covariate. For simplicity, we focus on three different substantive topics in this analysis. The first, which we label "attack," deals with concerns about the US attacking one's own country or society. The second, labeled "human rights," deals with posts about the implications of the Snowden episode for American credibility on issues related to freedom and human rights. The third, labeled "asylum," concerns news updates about Snowden's movements and whether or not he will be granted asylum and in which country.

First, we emphasize the role of the content covariates in handling the multiple source languages. For some reason, the Chinese version of Snowden's last name translated to "Snowdon", instead of "Snowden", while the Chinese version of Snowden's full name translated to "Edward Snowden". This was not the case in Arabic. Therefore, the Chinese examples were likely to use the word "Snowdon", in addition to "Snowden". "Snowdon" did not appear in the Arabic texts at all. Similarly, the Chinese encoding in Google Translate creates the word "quote" when a quotation mark appears. Therefore, many of the words in the Chinese corpus have "quote" attached to them, for example "quotsnowden" or "quotprism".

Of course, the analyst could go through and identify each of these mistakes and correct them, but this would be time consuming or impossible for larger tasks. By modeling the

[29]Note the similarity here to the multilingual models discussed in Section 3.2. While those models explicitly maintain models in two or more languages using external alignment information, here we are maintaining a model in only one language but allowing for limited residual variations from the original language. This provides a more parsimonious model structure and facilitates interpretation of the model results. Situations that call for an explicit representation of the topics within multiple languages would be better served by some of the alternatives discussed in Section 3.2.

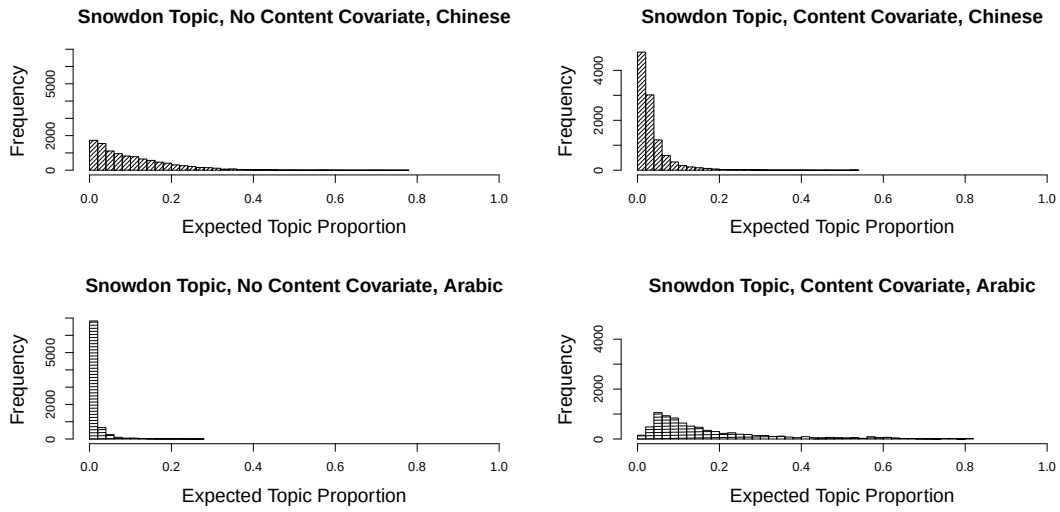

Figure 5: Histogram of topic proportions for the topic where the word "Snowdon" is most important. Without a content covariate, this topic is dominated by Chinese tweets and has very few Arabic tweets. With a content covariate, this topic mixes between Chinese and Arabic tweets.

fact that machine translation will make different mistakes in Chinese than in Arabic using a content covariate, we allow Chinese and Arabic tweets to talk about the same topic, while allowing the tweets from each language to use slightly modified versions of the vocabulary.

Consider the "Snowdon" mistake. For purposes of comparison, we ran a topic model that did not include a content covariate. Within this topic model, the word "Snowdon" pinned down its own topic. Because Snowdon was one of the words defining the topic, it was completely dominated by Chinese tweets; no Arabic tweets were estimated to have more than 0.1 of this topic (see Figure 5). However, this is a mistake. Chinese tweets translated to "Snowdon" are often discussing the same topic as Arabic tweeters using "Snowden". When we include the content covariate, "Snowdon" appears in a topic with "Snowden", and this topic is similarly distributed in Chinese and Arabic tweets. Had we failed to include a content covariate, we would have created a topic falsely associated with the Chinese tweets.

We now explore the results of the model and compare the full machine translation to the translation of the term-document matrix. To illustrate, we focus first on the two

topics related to the image of the United States in the eyes of Chinese and Arabic-language tweeters, namely the "attack" and "human rights" topics.

The "attack" topic, which discusses the U.S. "attacking" other countries, particularly focused on Snowden's allegations that the U.S. government hacked into Chinese government agencies and businesses. This topic contains words such as "China", "company", "attack", and "relationship". Many of the tweets question the U.S.-China bilateral relationship going forward. An example of the original and `translateR`-returned Google Translate translation of a tweet that contains a significant portion of this topic is shown in Figure 6. Note also that the translation captures the essence of the post.

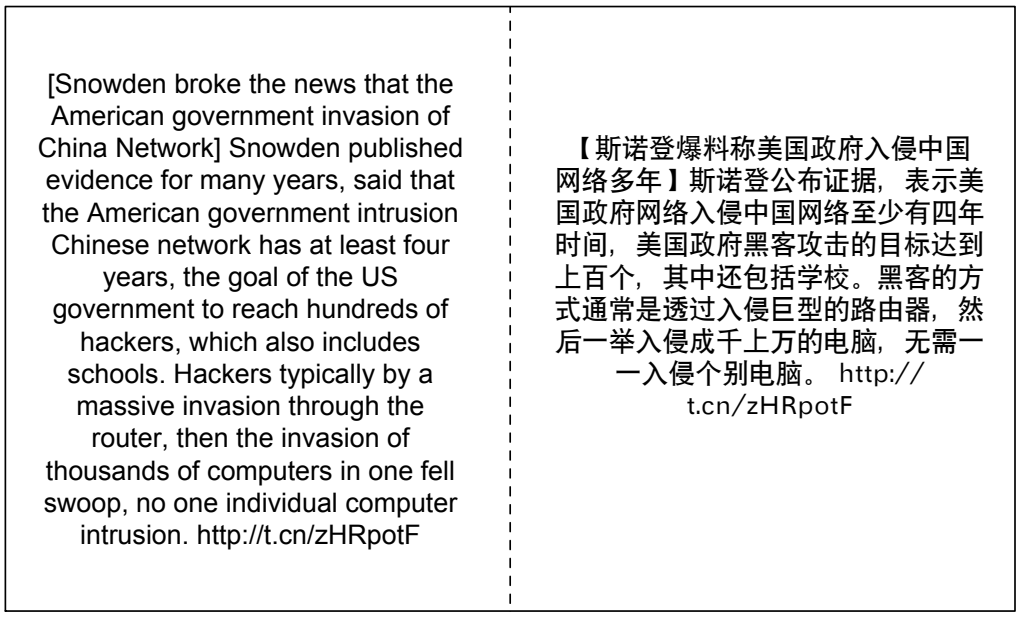

Figure 6: Example post for the attack topic.

The "human rights" topic discusses the U.S. record on human rights and whether the Snowden disclosures undermine this record. This topic contains words such as "violate", "freedom", "human", "right", and "traitor". Some of the posts also discuss whether

the U.S. is a hypocrite, violating U.S. citizens' human rights while also advocating for greater human rights protection abroad. An example of the original and `translateR` sourced-Google Translate translation of a tweet in this topic is shown in Figure 7.

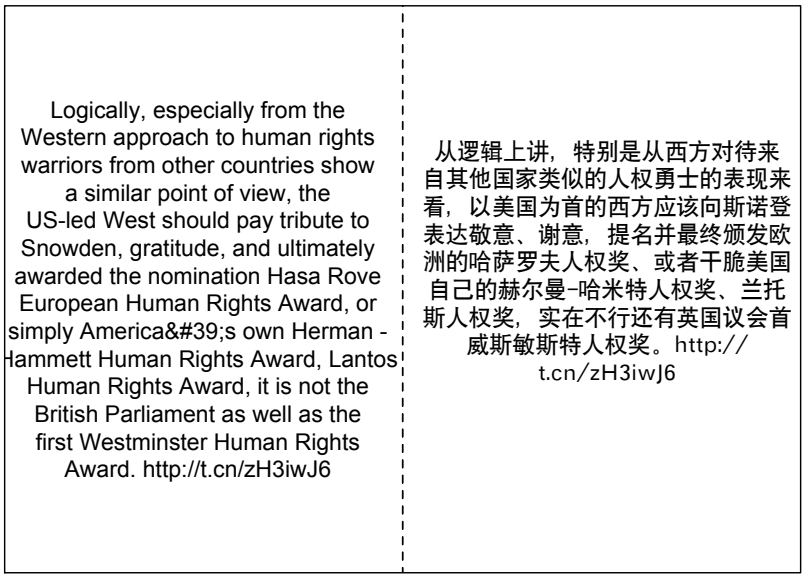

Figure 7: Example post for the human rights topic.

Given the dramatic cost difference between the full text and document-term matrix translations, it is useful to investigate the similarity of the resulting topic models. If the document-term matrix translation produces comparable results, it will clearly be preferable on cost alone. We investigate the similarity of the models for our two topics of interest, cautioning that congruence between the models for this case does not produce a general result.

We examine the alignment between models by comparing the topic-word distributions of all topics in both models. Because the two models use different vocabularies we identify the common terms and calculate the correlation between every pair of topics using the

overlapping words.[30] Some of the topics align quite clearly, including the two we have highlighted above. In Online Appendix F.4, we provide a visualization of the correlations between all topic pairs.

We explore the question of model alignment further by investigating how our aggregate inferences about the relative rates of topical prevalence would change under the different translation strategies. In Figure 8, we plot the three topics and their estimates under each of the two translation methods. Note that the three displayed topics - "Attack," "Human Rights," and "Asylum" - all have similar frequent terms and substantively similar estimates. For both the full text translation and the term-document matrix translation, the "Attack" and "Human Rights" topics are more associated with Chinese posts than with Arabic posts. At least in this case, two investigators using different translation methods might have reached similar substantive conclusions, namely that microbloggers in China seem very ready to condemn the U.S. government for hacking Chinese companies and the government and for "trampling" on human rights. This analysis is further explored in Online Appendix F, along with an overview of additional topics in the model and the technical details of the estimation.

Which topics are more associated with Arabic tweets? Arabic tweeters are more likely to be sharing news about the Snowden disclosures. The "Asylum" topic is associated with Arabic tweets and is related to speculation about where Edward Snowden will end up seeking asylum. This topic contains words such as "Ecuador", "Iceland", "shelter", "request", and "asylum". Arabic tweeters are much more likely to be sharing news, rather than opinions about the U.S. government's reputation.

These results begin to speak to our original interest in the ways that the reputation of the U.S. was damaged in the eyes of Chinese and Middle Eastern social media users during the Snowden incident. However, we also find that these topics were more prevalent within the Chinese corpus than the Arabic corpus. This is unlike other events where there are strong reactions to US intervention in the Middle East by Arabic twitter users (Jamal

[30]Specifically we construct a marginal estimate of the topic word distribution β by weighting the Chinese and Arabic specific version of the topics by their relative frequency in the corpus. We then take the intersection of the vocabulary between the two models and calculate the correlation between the distributions over those words for each topic pair.

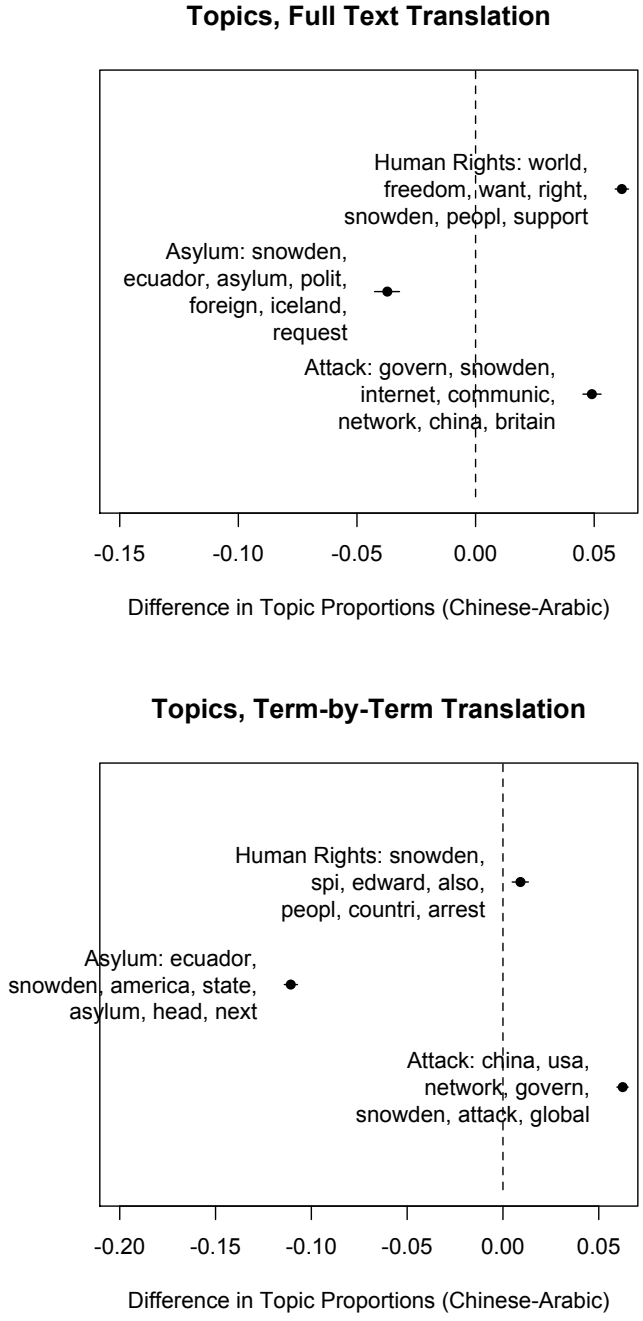

Figure 8: Topics related to U.S. reputation. The top plot is the estimation with full-text translation, the bottom plot is the estimation of these topics with DTM-translation. Both plots show the relationship between the topics and the Chinese and Arabic corpuses.

et al., nd). The Snowden disclosures seemed to affect Chinese perceptions of the U.S. more strongly; not only was there considerable outcry about US cyber-intervention in China, but the Snowden event generated discussion of how the US in fact opposes human rights and is less democratic than the US attempts to seem on the world stage. Perhaps, given the perception that US cyber activities targeted China, the Chinese response is consistent with previous work focusing on the Middle East. Using this type of workflow, scholars could examine reactions by many countries to other world events that the US is involved in.

5 Conclusion

The volume of textual data is growing rapidly throughout the world. The form of this textual data is no longer simply in the form of newspapers, books, etc., but also in social media and other internet based content that puts even fewer restrictions on the generation of textual data (e.g., Barberá, 2012). There is no sign that this trend will change. Even if a tiny fraction of this data is ultimately of interest to comparativists, they will need to understand a range of issues relevant to different languages that are actively being studied by scholars.

This paper introduces comparativists to a range of important topics in textual analysis. We walked through a variety of research questions that comparative politics scholars have been asking and answering with textual data, introduced the basics of textual processing with a focus on non-English texts. Next we discussed the managing and preprocessing of text from a multi-language perspective, including a brief discussion of new software such as `txtorg`, as well as a discussion of machine based translation where we introduce a new `R` package `translateR` that provides easy access to the Google Translate API. Next we briefly discussed techniques for text analysis, emphasizing the existing tools for multilingual text analysis. Finally we used the Structural Topic Model to provide two examples of how comparativists can use metadata to incorporate their knowledge of corpus structure into unsupervised learning, including a novel way to use the STM model when text has first been translated to a single language.

Future developments designed to address remaining challenges could proceed in a number of different directions. We are particularly interested in harnessing the ever increasing advances in automated translation with existing text analysis techniques. No doubt existing translation methods are imperfect; however, translation is an active research area in academia and industry which suggests that these system will continue to improve over time. An open question for social scientists is how to best leverage these developments for applied research. A critical part of this process is developing diagnostic tools for assessing the sensitivity of our analysis tools to translation error.

Finally, we plan to continue developing open-source software which brings the necessary tools for automated text analysis to the end user. The three software packages described here cover different portions of the text analysis workflow from processing of texts to estimating the model. We plan to continue refining these tools with comparative politics scholars in mind, while developing new software, including a browser based system for interactive topic model exploration.

References

Alfonseca, E., S. Bilac, and S. Pharies (2008). Decompounding query keywords from compounding languages. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 253–256. Association for Computational Linguistics.

Barberá, P. (2012). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. In *APSA 2012 Annual Meeting Paper*.

Baturo, A. and S. Mikhaylov (2013). Life of brian revisited: Assessing informational and non-informational leadership tools. *Political Science Research and Methods 1*(01), 139–157.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM 55*(4), 77–84.

Blei, D. M. and J. D. Lafferty (2007). A correlated topic model of science. *The Annals of Applied Statistics 1*(1), 17–35.

Boyd-Graber, J. and D. M. Blei (2009). Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 75–82. AUAI Press.

Boyd-Graber, J. and P. Resnik (2010). Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 45–55. Association for Computational Linguistics.

Brachman, J. (2009). *Global Jihadism*. New York: Routledge.

Brady, H. E. and D. Collier (2010). *Rethinking social inquiry: Diverse tools, shared standards*. Rowman & Littlefield.

Brown, P. F., J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin (1990). A statistical approach to machine translation. *Computational linguistics 16*(2), 79–85.

Brown, P. F., V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics 19*(2), 263–311.

Budge, I. (2001). *Mapping policy preferences: estimates for parties, electors, and governments, 1945-1998*, Volume 1. Oxford University Press.

Campbell, R. S. and J. W. Pennebaker (2003). The secret life of pronouns flexibility in writing style and physical health. *Psychological Science 14*(1), 60–65.

Catalinac, A. (2013). Pork to policy: The rise of national security in elections in japan.

Cheng, K.-S., G. H. Young, and K.-F. Wong (1999). A study on word-based and integral-bit chinese text compression algorithms. *Journal of the American Society for Information Science 50*(3), 218–228.

Chiozza, G. (2009). *Anti-Americanism and the American world order*. Baltimore: Johns Hopkins University Press.

Coscia, M. and V. Rios (2012). Knowing where and how criminal organizations operate using web content. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1412–1421. ACM.

Eggers, A. and A. Spirling (2011). Partisan convergence in executive-legislative interactions modeling debates in the house of commons, 1832–1915.

Farrell, H. and M. Finnemore (2013). End of hypocrisy: American foreign policy in the age of leaks, the. *Foreign Aff. 92*, 22.

Feinerer, I., K. Hornik, and D. Meyer (2008, March). Text mining infrastructure in r. *Journal of Statistical Software 25*(5), 1–54.

Fokkens, A., M. Van Erp, M. Postma, T. Pedersen, P. Vossen, and N. Freire (2013). Offspring from reproduction problems: What replication failure teaches us. In *ACL (1)*, pp. 1691–1701.

George, A. and A. Bennett (2005). *Case studies and theory development in the social sciences*. Mit Press.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America 101*(Suppl 1), 5228–5235.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis 18*(1), 1.

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

Habash, N. and J. Hu (2009). Improving arabic-chinese statistical machine translation using english as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 173–181. Association for Computational Linguistics.

Harman, D. (1991). How effective is suffixing? *JASIS 42*(1), 7–15.

Hollink, V., J. Kamps, C. Monz, and M. De Rijke (2004). Monolingual document retrieval for european languages. *Information retrieval 7*(1-2), 33–52.

Hu, Y., K. Zhai, V. Eidelman, and J. Boyd-Graber (2014). Polylingual tree-based topic models for translation domain adaptation. In *Association for Computational Linguistics*.

Hull, D. A. (1996). Stemming algorithms: a case study for detailed evaluation. *JASIS 47*(1), 70–84.

Jamal, A., R. O. Keohane, D. Romney, and D. Tingley (n.d.). Anti-americanism or anti-interventionism? evidence from the arabic twitter universe. *Perspectives on Politics forthcoming*.

Katzenstein, P. J. and R. O. Keohane (2007). Varieties of anti-americanism: A framework for analysis. In P. J. Katzenstein and R. O. Keohane (Eds.), *Anti-Americanisms in world politics*, pp. 9–38. Ithaca: Cornell University Press.

King, G., J. Pan, and M. E. Roberts (2013). How censorship in china allows government criticism but silences collective expression. *American Political Science Review 107*, 1–18.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Krovetz, R. J. (1995). Word-sense disambiguation for large text databases.

Laver, M., K. Benoit, and J. Garry (2003). Extracting policy positions from political texts using words as data. *American Political Science Review 97*(02), 311–331.

Lunde, K. (2009). *CJKV information processing*. O'Reilly Media, Inc.

Lynch, M. (2007). Anti-americanism in the arab world. In P. J. Katzenstein and R. O. Keohane (Eds.), *Anti-Americanisms in world politics*, pp. 196–224. Ithaca: Cornell University Press.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to information retrieval*, Volume 1. Cambridge University Press Cambridge.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.

McCants, W. (2006). Militant ideology atlas. Technical report, Combating Terrorism Center, U.S. Military Academy.

Miller, M. C. (2013). *Wronged by Empire: Post-Imperial Ideology and Foreign Policy in India and China*. Stanford University Press.

Mimno, D., H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pp. 880–889. Association for Computational Linguistics.

Mosteller, F. and D. L. Wallace (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association 58*(302), 275–309.

Nielsen, R. (2012). Jihadi radicalization of muslim clerics.

Nielsen, R. (2013). *The Lonely Jihadist: Weak Networks and the Radicalization of Muslim Clerics*. Ph. D. thesis, Harvard University. Ann Arbor: ProQuest/UMI. (Publication No. 3567018).

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics.

Paul, M., H. Yamamoto, E. Sumita, and S. Nakamura (2009). On the importance of pivot language selection for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 221–224. Association for Computational Linguistics.

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

Roberts, M. E., B. M. Stewart, and E. Airoldi (n.d.). A topic model for experimentation in the social sciences. *working*.

Roberts, M. E., B. M. Stewart, and D. Tingley (2014). *stm: R Package for Structural Topic Models*. R package version 0.6.21.

Roberts, M. E., B. M. Stewart, D. Tingley, and E. M. Airoldi (2013). The structural topic model and applied social science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. Gadarian, B. Albertson, and D. Rand (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science 58*(4), 1064 – 1082.

Rubin, B. (2002, December). The real roots of arab anti-americanism. *Foreign Affairs 81*(6), 73–85.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of*. Addison-Wesley.

Schonhardt-Bailey, C. (2006). *From the corn laws to free trade [electronic resource]: interests, ideas, and institutions in historical perspective*. The MIT Press.

Schrodt, P. A. and D. J. Gerner (1994). Validity assessment of a machine-coded event data set for the middle east, 1982-92. *American Journal of Political Science 38*(3), 825–854.

Slapin, J. B. and S.-O. Proksch (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science 52*(3), 705–722.

Stewart, B. M. and Y. M. Zhukov (2009). Use of force and civil–military relations in russia: an automated content analysis. *Small Wars & Insurgencies 20*(2), 319–343.

Stockmann, D. (2012). *Media commercialization and authoritarian rule in China*. Cambridge University Press.

Telhami, S. (2002). *The stakes: America and the Middle East*. Boulder, CO: Westview Press.

Tseng, H., P. Chang, G. Andrew, D. Jurafsky, and C. Manning (2005). A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Volume 171. Jeju Island, Korea.

Utiyama, M. and H. Isahara (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *HLT-NAACL*, pp. 484–491.

Van Atteveldt, W., J. Kleinnijenhuis, and N. Ruigrok (2008). Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from dutch newspaper articles. *Political Analysis 16*(4), 428–446.

Volkens, A., P. Lehmann, N. Merz, S. Regel, A. Werner, O. Lcewell, and H. Schultze (2013). The manifesto data collection. *Manifesto Project (MRG/CMP/MARPOR), Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB)*.

Zhao, B. and E. P. Xing (2006). Bitam: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 969–976. Association for Computational Linguistics.